Royal Meteorological Society

Monte Carlo Bayesian inference on a statistical model of sub-gridcolumn moisture variability using high-resolution cloud observations. Part 1: Method

Peter M. Norris[a,b]* and Arlindo M. da Silva[b]

[a] *Goddard Earth Sciences Technology and Research, University Space Research Association, Columbia, MD, USA*
[b] *Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA*

*Correspondence to: P. M. Norris, Global Modeling and Assimilation Office, NASA/GSFC, Code 610.1, Greenbelt, MD 20771, USA. E-mail: peter.m.norris@nasa.gov

A method is presented to constrain a statistical model of sub-gridcolumn moisture variability using high-resolution satellite cloud data. The method can be used for large-scale model parameter estimation or cloud data assimilation. The gridcolumn model includes assumed probability density function (PDF) intra-layer horizontal variability and a copula-based inter-layer correlation model. The observables used in the current study are Moderate Resolution Imaging Spectroradiometer (MODIS) cloud-top pressure, brightness temperature and cloud optical thickness, but the method should be extensible to direct cloudy radiance assimilation for a small number of channels. The algorithm is a form of Bayesian inference with a Markov chain Monte Carlo (MCMC) approach to characterizing the posterior distribution. This approach is especially useful in cases where the background state is clear but cloudy observations exist. In traditional linearized data assimilation methods, a subsaturated background cannot produce clouds via any infinitesimal equilibrium perturbation, but the Monte Carlo approach is not gradient-based and allows jumps into regions of non-zero cloud probability. The current study uses a skewed-triangle distribution for layer moisture. The article also includes a discussion of the Metropolis and multiple-try Metropolis versions of MCMC.

Key Words: cloud data assimilation; statistical cloud parametrizations; Bayesian inference; Markov chain Monte Carlo

Received 18 April 2016; Accepted 19 May 2016; Published online in Wiley Online Library 27 July 2016

1. Introduction

This article addresses the topic of cloud data assimilation (CDA) into global circulation models (GCMs). CDA presents many challenges, including the following.

(1) Cloud properties and the underlying moisture field for which they are partial markers often have very significant variability on scales smaller than typical GCM gridbox sizes.
(2) Cloud-affected observables vary nonlinearly with gridcolumn profiles of temperature and moisture. In particular, no infinitesimal *equilibrium* perturbation to moisture for a subsaturated background state can produce cloudiness. This makes dealing with cloudy observations and a clear background an especially difficult problem.
(3) The forward modelling of cloud-affected radiances is complicated and computationally expensive.
(4) Due to the complexity of cloud macro- and microphysical processes and their coupling with convection, turbulence and radiative transfer, the generation of a good background cloud state is itself a very difficult task. This makes attempts at cloud data assimilation from this background state all the more difficult, since most data assimilation methods work best when the background is close to the observed state. In particular, small errors in synoptic system location are much more serious for cloud data assimilation than for the variational constraint of smoother fields such as temperature.
(5) Clouds are directly coupled to the flow field and thermodynamic state and assimilation of cloud data therefore requires building the full state that supports them.

Because of these and other difficulties, much of the available cloud-affected satellite data are currently discarded. These discarded observations comprise a significant fraction of all satellite data. This is a major problem, because cloud-affected radiances carry very important information about the underlying moisture field and ignoring this information potentially introduces a significant bias into the moisture analysis.

Our intention in this article is not to attack the entire CDA problem, but to explore a non-traditional Monte Carlo Bayesian approach to cloud analysis – to improving the background moisture state and its subgrid-scale variability, so that it is much more compatible with the cloud observations provided. Such an approach is designed to be a type of 'cloud relocator' or 'cloud bias-corrector', able to provide more traditional variational cloud data assimilation methods (see Bauer *et al.*, 2011) with a background state that yields forward-modelled cloudy radiances with much smaller biases with respect to the observations.

Bauer *et al.* (2011) review current efforts under way at global numerical weather prediction (NWP) centres to assimilate cloud and precipitation data. These efforts largely involve variational (3D- or 4D-Var) assimilation of cloud-affected radiances (or cloud property retrievals). Some centres, such as the UK Met Office and Météo-France, include a 1D-Var cloud preprocessing step – to facilitate quality control, to produce a cloud-observation-consistent pseudo-relative-humidity for assimilation by the full CDA system or to constrain unknown cloud-related parameters of the system, which are subsequently held fixed in the next forecast cycle. The authors' earlier work (Norris and da Silva, 2007) was of this nature, using International Satellite Cloud Climatology Project (ISSCP) and Special Sensor Microwave Imager (SSM/I) cloud-property retrievals to constrain the empirical parameters of a simple GCM cloud parametrization. The approach presented here has similarities with these 1D-Var approaches, but with the following significant differences.

(1) The approach is fully Bayesian, seeking to characterize the *a posteriori* probability density function (PDF) of control parameters. In particular, the approach makes no assumptions of Gaussianity, which are potentially dangerous in treating the often skewed moisture distributions found in nature, e.g. in the marine boundary layer.

(2) The approach is fully nonlinear and makes no use of gradient or adjoint information. In particular, the Monte Carlo optimization method can jump out of zero-sensitivity regions, such as subsaturated gridcolumns in the presence of cloud observations.

(3) The approach makes use of a detailed sub-gridcolumn statistical model, including horizontal moisture variability within each gridbox and its vertical coupling between layers. In particular, cloud-affected observations within the gridcolumn footprint are used to constrain second and third statistical moments within the gridcolumn, not just mean properties. This partly mitigates the representativeness errors associated with the typical mismatch between satellite pixel footprints and the gridcolumn footprint.

The multivariate aspect of coupling clouds to the dynamic and thermodynamic state (item (5) earlier) deserves much attention in a fully cycled cloud data assimilation system, but is not considered in this article. Coupling of our Monte Carlo Bayesian algorithm to the ensemble/variational schemes currently implemented in the National Aeronautics and Space Administration (NASA) Global Modeling and Assimilation Office (GMAO) Goddard Earth Observing System Model Version 5 (GEOS-5) data assimilation system is beyond the scope of this article.

This article is Part 1 of a two-part series. In section 2, a detailed description of our method is presented, since a number of aspects of the method are non-traditional in the CDA field and therefore warrant careful explanation. Then, in section 3, an application of the method to some simple single-layer case studies is illustrated and analyzed.

In the sequel, Part 2, the method is applied in a realistic multi-layer setting and its performance in a number of case studies and sensitivity tests is discussed.

2. Description of method

2.1. Overview

This article presents a type of Bayesian parameter estimation, somewhat akin to 1D-Var approaches, in which the fundamental unit of estimation is a single GCM-gridcolumn. However, unlike 1D-Var, the gridcolumn is not simply a profile of layer-mean state variables, but a more comprehensive statistical model designed to capture more realistic subgrid-scale horizontal and vertical variability within the gridcolumn. This model includes a PDF of subgrid total moisture for each layer and a suitable coupling of those PDFs in the vertical, e.g. using a Gaussian copula, as in Norris *et al.* (2008).

Let $\boldsymbol{\alpha}$ be a control vector of the parameters that describe the gridcolumn model: $\boldsymbol{\alpha}$ should include a set of parameters specifying each layer's moisture PDF (not just its mean) and may also include the vertical coupling between the layer PDFs, e.g. via one or more vertical decorrelation length-scales associated with inter-layer correlation.

We have some assumed prior knowledge of $\boldsymbol{\alpha}$ given by a prior PDF $p(\boldsymbol{\alpha})$, and we wish to explore how this knowledge is modified by a vector of observations of the gridcolumn, $\boldsymbol{y}$. In other words, we wish to estimate the posterior PDF $p(\boldsymbol{\alpha}|\boldsymbol{y})$. In our case, $\boldsymbol{y}$ contains cloud retrievals for satellite instrument fields of view (FOVs) falling within the gridcolumn, e.g. cloud-top temperature and cloud optical thickness. There are usually many such FOVs within the gridcolumn footprint, so these observations should contain information useful for constraining not only the gridcolumn mean state but also its internal spatial variability.

According to Bayes' Theorem,

$$p(\boldsymbol{\alpha}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\alpha})\, p(\boldsymbol{\alpha}), \tag{1}$$

with the constant of proportionality $1/p(\boldsymbol{y})$ being dependent only on the observations. The first term on the right, $p(\boldsymbol{y}|\boldsymbol{\alpha})$, is called the likelihood, the probability of observing $\boldsymbol{y}$ given a parameter state $\boldsymbol{\alpha}$. There are usually a large multiplicity of parameter states $\boldsymbol{\alpha}$ that have some likelihood of yielding the observations. Our goal is to explore $p(\boldsymbol{\alpha}|\boldsymbol{y})$ to quantify both its mode(s), which specifies the most probable $\boldsymbol{\alpha}$ given the observations, as well as some measure of its spread, which indicates the magnitude of the error associated with the modal $\boldsymbol{\alpha}$ estimate.

For reasons to be discussed below, we will use a type of Markov chain Monte Carlo (MCMC) method to characterize $p(\boldsymbol{\alpha}|\boldsymbol{y})$. This method makes quasi-random jumps around parameter space, such that, as the number of jumps becomes large, the collection of sampled $\boldsymbol{\alpha}$ is consistent with samples drawn from the target PDF $p(\boldsymbol{\alpha}|\boldsymbol{y})$. (Note that, while sampling $\boldsymbol{\alpha}$ space, the proportionality term $p(\boldsymbol{y})$ is invariant and need not be included.)

The above approach has a number of benefits, at least some of them specific to highly nonlinear problems such as CDA.

(1) It is not based on gradient or tangent-linear sensitivity, but instead makes exploratory jumps in parameter space. This potentially allows the method to jump out of regions of zero likelihood, such as when cloudy observations fall within a wholly subsaturated gridcolumn. It also allows sampling away from local but non-global probability maxima, with the potential to find the global maximum. Of course, non-gradient methods come with an increased computational cost, which is often a very important concern.

(2) Most common variational approaches assume Gaussian likelihood and prior models and are therefore expressed in terms of least-squares cost functions. While some type of transformation to a Gaussian likelihood may be possible, the Bayesian approach above does not require this and so is somewhat more flexible.

(3) The MCMC approach characterizes the posterior PDF in general, not just the mode. Again, this comes at a cost, but the advantage is that robust error estimates of the parameters are available from the method.

(4) The MCMC approach implicitly produces a set of samples of parameter space consistent with the *a posteriori* parameter PDF. This set can potentially be used to produce a sampling of the new prior for the next analysis time via an ensemble forecast in $\boldsymbol{\alpha}$.

(5) In general, there will be many different candidate statistical models for the gridcolumn: for example with different forms for the layer moisture PDFs or the inter-layer correlation (or 'cloud overlap'). Conveniently, the Bayesian context can be used to decide which model is more consistent with the observations. For a model $\mathcal{M}$ with parameters $\boldsymbol{\alpha}$, the probability of the observations is $p(\boldsymbol{y}|\mathcal{M}) = \int p(\boldsymbol{y}|\mathcal{M},\boldsymbol{\alpha})\,p(\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha}$ and can be approximated as the average of the likelihood term sampled over the prior distribution of $\boldsymbol{\alpha}$. This can easily be compared with $p(\boldsymbol{y}|\mathcal{M}')$ of an alternate model $\mathcal{M}'$ in $\boldsymbol{\alpha}'$ to decide which model is better. (The prior samplings at a particular time can readily be supplied from the *a posteriori* sampling at the previous analysis, as in (4) above.)

As noted already, the above set of benefits come with an additional computational cost and a cost-benefit analysis may not always favour the chosen approach for a given problem. If only a maximum *a posteriori* estimate is required (i.e. the mode of $p(\boldsymbol{\alpha}|\boldsymbol{y})$), not a full characterization of $p(\boldsymbol{\alpha}|\boldsymbol{y})$, then techniques such as simulated annealing (see e.g. Andrieu *et al.*, 2003) may offer significant time savings. Nevertheless, the computational throughput for the MCMC-based CDA system described here is quite manageable. Actual timings are given in Appendix B2.

2.2. *Gridcolumn statistical model*

The 'gridcolumn statistical model' (GCSM) encapsulates and parametrizes the horizontal moisture variability within each layer of a GCM-gridcolumn and its vertical correlation between layers. We employ this GCSM construct as the fundamental unit of our analysis, because there is well-known and significant variability in moisture/cloud fields at scales below the typical GCM grid-scale and because high-resolution satellite cloud observations have the potential to help constrain this variability.

Much of the background for the GCSM comes from Norris *et al.* (2008, hereafter N08). The moisture variable we use is the total saturation ratio, $S \equiv q_\mathrm{t}/q_\mathrm{s}(T)$, the ratio of the *in situ* total moisture content (vapour plus cloud condensate) to the saturation vapour content at the *in situ* temperature T. We model the intra-layer variability in S, but neglect any explicit intra-layer temperature variability, using the layer mean temperature $\bar{T}$ whenever an explicit temperature is required. N08 provide a solid justification for this approach.

The first key component of the GCSM is the PDF $p_{S_k}(S)$ of moisture variability for each layer k of the gridcolumn. For sufficiently thin layers, this quantifies the intra-layer *horizontal* variability of S_k. In general, each layer can have its own unique form for p_{S_k}. In practice, it is simpler to pick a common parametric form $p_S(S;\,\boldsymbol{\nu}_k)$, where $\boldsymbol{\nu}_k$ is the layer parameter vector, fixed in dimension but varying in value between layers. This is acceptable so long as the form is general enough to encompass realistic variability at different heights in the atmosphere, at different latitudes and under different synoptic conditions. This argues for flexibility to control at least the mean, variance and skewness of the PDF, or at least three parameters per layer.

N08 use a generalized extreme value (GEV) distribution for p_S. In the current article, we use a simpler skewed triangle distribution, denoted $p_\triangle[S;\,\boldsymbol{\nu} \equiv (S_\mathrm{L}, S_*, S_\mathrm{H})]$ and illustrated in Figure 1. This PDF is non-zero on $(S_\mathrm{L}, S_\mathrm{H})$, rising linearly from zero at S_L to a mode at S_* and then falling linearly to zero again at S_H. As such, it is a simple three-parameter bounded PDF that retains skewness. We prefer this ability to represent skewness because of the ubiquity of skewed moisture PDFs in the boundary layer and in convective regimes. A bounded PDF is convenient to avoid unphysical negative S values or unrealistically large S values (see sections 2.3.1 and 2.8). Appendix A has a fuller description of the properties of $p_\triangle$.

The second key component of the GCSM is the coupling or correlation of moisture PDFs among the layers. This is a generalization of the familiar 'cloud overlap' concept to the vertical correlation of the layer moisture PDFs themselves, not just the cloud occurrence. In the current article, the vertical stack of triangular PDFs is coupled using a Gaussian copula, as described in N08.

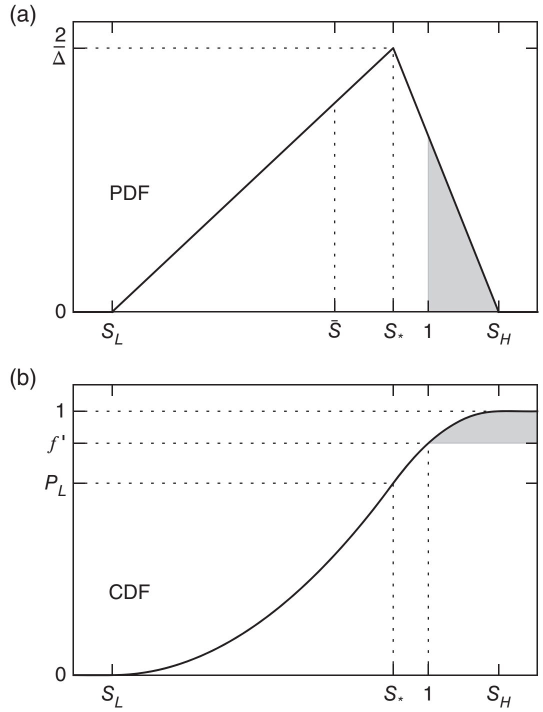

Figure 1. Example of (a) a skewed triangle PDF in S and (b) its CDF. The PDF has width $\Delta \equiv S_\mathrm{H} - S_\mathrm{L}$. The fraction of probability to the left of the mode is $P_\mathrm{L} = (S_* - S_\mathrm{L})/\Delta$. The shaded area of the PDF is the cloud fraction $f = 1 - f'$ (under the bulk assumption). See Appendix A for further details.

Formally speaking, the copula of a set a random variables is the joint cumulative distribution function (CDF) of the ranks of the variables within their respective univariate margins. In our context, it is a means of producing a full multi-layer distribution of S with the specified layer marginal PDFs p_{S_k}. As in N08, we model the copula of S with the well-known and convenient Gaussian copula, parametrized by an inter-layer correlation matrix $\mathbf{C}$. The resultant multi-layer S distribution is *not* Gaussian, but it does collapse to a multivariate Gaussian with correlation $\mathbf{C}$ if the layer marginals are themselves Gaussian. N08 provides solid justification for the use a Gaussian copula, demonstrating its ability to model complex sub-GCM-gridcolumn cloud fields while retaining the convenience of parametric marginal PDFs.

It is a simple matter to generate random samples from the Gaussian copula/triangular margin combination presented above (see N08 and Appendix A1). We call this subcolumn generation, in the sense that each such random sample is a vector of layer S values that can be used to specify a full vertical profile (or 'subcolumn') of vapour content, cloudiness and condensate at a random horizontal location within the gridcolumn footprint. The generation of an ensemble of such subcolumns permits Monte Carlo integration over the subgrid-scale variability within a GCM-gridcolumn. This approach is basically necessary due to the intractability/computational expense of analytical/other numerical gridcolumn averages, for all but the most simple forward operators. A similar approach is used by the McICA radiative transfer calculations of Pincus *et al.* (2003).

We make the so-called bulk assumption, namely that the condensate amount is the total water in excess of saturation. Then layer k of a subcolumn is either cloudy for $S_k > 1$ or clear for $S_k \leq 1$ and the total, vapour and condensate contents are given by

$$\begin{aligned} q_{\mathrm{t}k} &= S_k\, q_{\mathrm{s}k}, \\ q_{\mathrm{v}k} &= \min(S_k, 1)\, q_{\mathrm{s}k}, \\ q_{\mathrm{c}k} &= q_{\mathrm{t}k} - q_{\mathrm{v}k} = \max(S_k - 1, 0)\, q_{\mathrm{s}k}, \end{aligned} \tag{2}$$

respectively (where $q_{sk} = q_s(\bar{T}_k)$ is used, as explained earlier). The bulk assumption, together with the phase split details discussed later, provides a simple framework for the investigation of subgrid-scale moisture variability in terms of total (vapour + condensate) moisture content (although it is admittedly simplistic for mixed-phase and ice clouds, as we will discuss later and hope to improve upon in future work.)

The above subcolumn generation strategy is a type of independent column approximation (ICA) subcolumn generation, in the sense that our GCSM does not contain any specification of horizontal correlation, just horizontal PDFs of S coupled only in the vertical. Wind *et al.* (2013) provide a treatment of horizontal spatial coherence for subcolumn generation that becomes essential in the context of cloud retrieval simulations. However, inclusion of horizontal coherence is a computationally expensive proposition that is inconsistent with the ICA approach adopted in the GEOS-5 GCM and is therefore not considered in this study.

To complete the specification of the GCSM, the correlation matrix **C** must be specified. This is somewhat akin to the choice of cloud overlap parametrization, but more general, since it actually specifies the full relationship between S values (clear or cloudy) at different heights. Currently we assume a single decorrelation 'length'-scale of $L = 100$ hPa and use the well-known second-order autoregressive (SOAR) correlation function (e.g. Daley, 1991, p117):

$$\mathrm{C}_{kk'} = (1 + \xi_{kk'}) \exp(-\xi_{kk'}), \qquad \xi_{kk'} = |p_k - p_{k'}|/L, \qquad (3)$$

where p_k and $p_{k'}$ are the midpoint pressures of arbitrary layers within the gridcolumn. Many other correlation functions could be used, including ones that use more than one length-scale, as in Gaspari *et al.* (2006). The value $L = 100$ hPa was chosen, somewhat arbitrarily, to be approximately consistent with a 1 km low-level correlation length. This value is a reasonable global average estimate for the length-scale associated with the correlation of ranks of total *condensate* amount (see figure 3 of Oreopoulos and Norris, 2011), whereas our L is the length-scale associated with the correlation of ranks of total *water* (vapour + condensate) amount. Nevertheless, the two will be very similar for correlations between overcast cloud layers, where the vapour is saturated. The reader is referred to Part 2 (Table 3 and associated text) for discussion of the sensitivity to L and additional ways to specify L or infer it statistically, or to modify this constant default length-scale effectively in the presence of decorrelating atmospheric features, such as the buoyancy-inhibiting temperature inversions capping planetary boundary-layer clouds. A fuller study of the specification of **C** and L is planned.

Finally, note that the above GCSM is simply the one we have chosen for the preliminary evaluation of our CDA system. Many details can be altered in future studies. These include the form of the layer PDFs, the specification of **C**, for example using multiple length-scales, and whether to use the Gaussian copula or some other sort of cloud/condensate overlap model. One advantage of our CDA system is the ease with which such changes can be made, because of the abstraction of the subcolumn generator from other parts of the CDA system.

2.3. *Control vector and prior*

The control vector $\boldsymbol{\alpha}$ of the Bayesian inference (1) must include whatever parameters of the GCSM are allowed to vary in response to observations. Currently, $\boldsymbol{\alpha}$ is comprised of the PDF parameters $\boldsymbol{\nu}_k$ for each layer, hence allowing the observations to constrain the mean, variance and skewness of the layer moisture distributions. We do not include the gridbox mean temperature $\bar{T}$ profile, but hold it fixed during the CDA process. There is nothing in the algorithm that prevents the inclusion of $\bar{T}$, but because temperature is perhaps the best constrained parameter by the current meteorological observing system, we have decided in the current study to focus on correcting errors in the moisture field. Likewise, we do not include the currently fixed vertical decorrelation scale L, although this is very possible and will be tried in future studies.

To proceed, we first need to specify the prior $p(\boldsymbol{\alpha})$ at the time t_{a} of the analysis (i.e. inference). In a cycling CDA system, this prior should somehow be generated from a knowledge of the posterior PDF $p(\boldsymbol{\alpha}|\boldsymbol{y})$ at the *previous* analysis time and an idea of how to evolve the control vector $\boldsymbol{\alpha}$ in time. Since the time evolution of the moisture field is nonlinear, especially in the presence of cloud processes, the most thorough approach would be to draw a sufficiently large ensemble of $\boldsymbol{\alpha}$ samples from the previous posterior, evolve each sample forward in time with a prognostic PDF cloud parametrization, for example, following the lead of Tompkins (2002), and then form the prior at the new time from the ensemble of evolved $\boldsymbol{\alpha}$ samples.

In the current article and its sequel, Part 2, we step back from this thorough approach in several major ways.

(1) Firstly, we do not currently have a cloud parametrization that can prognosticate the $\boldsymbol{\alpha}$ control vector (i.e. the moisture PDFs) directly within the GEOS-5 GCM. Instead, we can *potentially* evolve $\boldsymbol{\alpha}$ implicitly by using the current GEOS-5 parametrizations, which prognosticate the gridbox mean vapour and condensate and the gridbox cloud fraction and then transform backwards and forwards to the PDF parameter space $\boldsymbol{\nu}_k$. Transformation from the GEOS-5 mean state to $\boldsymbol{\alpha}$ space is covered in section 2.3.1. The reverse transformation is simple to obtain from Appendix A.

(2) Secondly, we do not currently use an ensemble forecast of $\boldsymbol{\alpha}$ from the previous posterior PDF $p(\boldsymbol{\alpha}|\boldsymbol{y})$ to form the new prior at t_{a}. This is not because it cannot be done, nor fundamentally because of limitation (1) above, but because we have not currently integrated our CDA work with recent ensemble forecast advances at the GMAO. Currently, instead of the ensemble approach, we simply approximate the 'location' of the new prior by a transformation to $\boldsymbol{\alpha}$ space of the mean GEOS-5 gridcolumn state at t_{a} and use assumptions to specify its form (i.e. higher moments). We will be more precise in sections 2.3.1–2.3.3 below.

(3) Finally, currently we do not perform a cycling analysis at all, but rather a series of three-hourly cloud analyses that are 'independent' of each other in the following sense: the prior at each analysis time is derived *not* from the previous CDA posterior but from the state at t_{a} of a so-called background forecast run of the GEOS-5 model within the regular GMAO meteorological analysis cycle, which currently has no CDA input. This 'background' for our cloud analyses is thus not affected by the CDA at all, but is simply the source of the mean gridcolumn states used to produced the prior and initial condition for each cloud analysis. Details are covered in sections 2.3.1 and 2.4 below. The current non-cycling system is simply the scope of this article and its sequel, which focus on the cloud analyses, not a fundamental limitation imposed by (1) or (2) above. Future work will study a cycling CDA system with and without limitations (1) and (2).

2.3.1. *The background $\boldsymbol{\alpha}$ state*

A background run of the GMAO GEOS-5 GCM within its regular meteorological analysis cycle provides forecasts of the mean vapour and condensate contents, $\bar{q}_{\mathrm{v}}$ and $\bar{q}_{\mathrm{c}}$, in each gridbox, together with the cloud fraction f. Ideally, the GCM would provide estimates of the background $\boldsymbol{\alpha}$ (i.e. PDF parameters) directly, through a prognostic triangular PDF cloud parametrization. However, until such a prognostic PDF cloud parametrization is available in GEOS-5, a method is required to convert from $(\bar{q}_{\mathrm{v}}, \bar{q}_{\mathrm{c}}, f)$ to the triangular PDF parameters $\boldsymbol{\nu} = (S_{\mathrm{L}}, S_*, S_{\mathrm{H}})$. This method is described below. The collection of these $\boldsymbol{\nu}_k$ for each layer is called the 'background' control vector $\boldsymbol{\alpha}$ and is used by the CDA system in two ways: (i) to locate the prior in $\boldsymbol{\alpha}$ space, as described in section 2.3.3, and (ii) as the initial condition for sampling of the posterior $p(\boldsymbol{\alpha}|\boldsymbol{y})$, i.e. as the first element of the MCMC chain (section 2.7).

In general, one would expect that three constraints $(\bar{q}_v, \bar{q}_c, f)$ would be sufficient to solve for a unique triangle PDF specified by three parameters (S_L, S_*, S_H). It turns out that the situation is far more complex. Before examining this in detail, we switch from (S_L, S_*, S_H) to an equivalent but more convenient specification of the triangle PDF using $(\bar{S}, \Delta, P_L)$, where $\Delta \equiv S_H - S_L > 0$ is the base of the triangle, $P_L = (S_* - S_L)/\Delta$ is the fraction of probability to the left of the mode and $\bar{S}$ is the mean (see Appendix A, sections A1 and A2).

There are three cases to consider. Firstly, if $\bar{q}_c$ is zero, this is taken as sufficient evidence of a clear gridbox, regardless of f, which is considered to be less reliable. In this case, the entire total water triangular PDF must fall below saturation ($S = 1$). Specifically, we seek a triangle that falls wholly *within* $S \in [0, 1]$ and for which the mean $\bar{S}$ is equal to the specified $\bar{q}_v/q_s(\bar{T}) < 1$. These conditions are not sufficient to determine a unique skewed triangle PDF. Instead we impose a symmetric triangle ($P_L = 0.5$) centred on $\bar{S}$ with a nominal width $\Delta_0 = 0.4$. If this triangle does not fit wholly within $S \in [0, 1]$, we make adjustments to it, as detailed in section A5.

Secondly, if $\bar{q}_v = q_s$, this is taken as sufficient evidence of an overcast gridbox, regardless of f, which again is considered less reliable. [Note that this precise equality exists because, under the bulk assumption, any water in excess of saturation is condensate. Therefore q_v is everywhere $\leq q_s$ and so the gridbox mean $\bar{q}_v$ can only be equal to q_s if the gridbox is entirely saturated. When preprocessing the GEOS-5 background state, any value of $\bar{q}_v$ in excess of q_s is clipped to q_s.] In this overcast case, the entire total water triangular PDF must fall above saturation ($S = 1$). Specifically, we seek a triangle that falls wholly within $S \in [1, S_{\max}]$, where $S_{\max}$ is some reasonable upper bound on allowable total saturation ratio S (see section 2.8) and for which the mean $\bar{S}$ is equal to the specified $\bar{q}_t/q_s = 1 + \bar{q}_c/q_s > 1$. These conditions are not sufficient to determine a unique skewed triangle PDF. Instead we again impose a symmetric triangle ($P_L = 0.5$) centred on $\bar{S}$ with a nominal width $\Delta_0 = 0.4$. If this triangle does not fit wholly within $S \in [1, S_{\max}]$, we make adjustments to it, as detailed in section A5. Note that the use of an upper bound in total saturation ratio $S_{\max}$ is necessary to avoid unphysically large values of condensate q_c that may sometimes occur during the minimization. The bulk condensate assumption (2) simply does not contain the necessary precipitation microphysics to self-limit excessive q_c values. The specification of $S_{\max}$ is discussed in section 2.8.

Finally, in all other cases, the gridbox is partially cloudy and we diagnose a skewed triangle PDF that straddles over the saturation point ($S = 1$) from $(\bar{q}_v, \bar{q}_c, f)$, as detailed in Appendix A4. This turns out to be a non-trivial procedure, because a valid triangular solution is only available for a relatively narrow range of f values given $\bar{q}_v$ and $\bar{q}_c$. An f outside this range suggests that the underlying background PDF is not well approximated by a triangle. In that case, f is adjusted to provide a reasonable triangular background PDF (see section A4). In all cases, whether clear, overcast or partially cloudy, the mean $\bar{S}$ from the GEOS-5 background state is *always* preserved by the derived background-state triangle PDF.

2.3.2. *Choice of control parameters*

Each GCM layer below a nominal tropopause has a triangular PDF in total saturation ratio S. (PDFs could also have been specified above the tropopause, but this was not done to save on computational expense, since clouds generally do not occur there.) As per the previous section, reasonable control parameters are $(\bar{S}, \Delta, P_L)$, which are, respectively, the mean, the base of the triangle (representing PDF spread) and a non-dimensional skewness parameter $P_L \in (0, 1)$. $\bar{S}$ is allowed to vary in the interval $(0, S_{\max})$, where $S_{\max}$ is a maximum S in the range 1.1–1.4, depending on the assumed phase, as discussed in section 2.8. P_L is the fraction of probability to the left of the mode. In practice, we constrain P_L to a reduced range 0.1–0.9, representing extreme positive and negative skewnesses, respectively.

We do not use the raw base width $\Delta = S_H - S_L$, but a scaled version, $\beta \equiv \Delta/\Delta_{\max}$, as a control parameter. Here $\Delta_{\max}$ is the maximum possible base width that avoids the triangle end-points crossing out of physical/reasonable bounds, namely either $S_L < 0$ or $S_H > S_{\max}$, and is given by

$$\Delta_{\max} = 3 \min\left(\frac{\bar{S}}{1 + P_L}, \frac{S_{\max} - \bar{S}}{1 + P_H}\right). \tag{4}$$

Use of β rather than Δ provides a control parameter that is normalized on the invariant range $[0, 1]$, which is convenient for our algorithm and also provides a control variable with a smaller dynamic range, since Δ itself can become very small in the dry upper atmosphere.

So, with this modification, the collection of control parameters $(\bar{S}, \beta, P_L)$ for all layers in a gridcolumn (below the nominal tropopause) makes up the control vector $\boldsymbol{\alpha}$ used in our Bayesian estimation.

2.3.3. *Prior PDFs*

In the absence of prior knowledge to the contrary, we currently assume the prior PDFs of β and P_L to be independent uniform distributions on $[0, 1]$ and $[0.1, 0.9]$, respectively, for each tropospheric layer. We therefore impose no preferred values or prior correlations between layers for these variables. Clearly, in a future cycling CDA system, there is the potential of using *posteriori* knowledge of these variables from the previous CDA analysis to inform the prior for the current analysis, as discussed above in the introductory remarks of section 2.3.

For $\bar{S}$, we use a multivariate Gaussian prior centred on the vector of background tropospheric layer means and with covariance matrix

$$\Sigma_{kk'} = \sigma_{\bar{S}}^2\, r_k\, r_{k'}\, C_{kk'}, \tag{5}$$

where $\sigma_{\bar{S}} = 0.1$ is a nominal standard deviation for $\bar{S}$ and $C_{kk'}$ is the inter-layer correlation matrix of (3). The factor $r_k \equiv (p_k - p_{\text{lim}})/(p_{\text{ramp}} - p_{\text{lim}})$, restricted to $[0, 1]$, acts to ramp the standard deviations linearly to zero over a small pressure interval near the tropopause. Currently, $p_{\text{lim}} = 50$ hPa and $p_{\text{ramp}} = 100$ hPa.

The use of a Gaussian prior for $\bar{S}$ has some justification in the work of Dee and da Silva (2003), who find that relative humidity is a preferred and more Gaussian moisture control variable than specific humidity. However, in some sense, we are also simply following the common practice of assuming a Gaussian as a first step in algorithm development. The constant value of $\sigma_{\bar{S}} = 0.1$ is also somewhat arbitrary and motivated by the fact that typical $\bar{S}$ values are in the range $[0, S_{\max}] \approx [0, 1]$. We also assume no prior correlation between $\bar{S}$ and either of β or P_L.

Clearly the implications of these assumptions need to be investigated thoroughly. At this stage, we have specified the simplest prior PDF for the parameters and focused on other aspects of the MCMC algorithm. In the future, the prior will take on a form of its own based on ensemble evolution of the posterior in a cycling system, as discussed earlier in section 2.3.

2.4. *Observations*

Our CDA method produces an analyzed model state at nominal time t_a using cloud data within a 3 h window centred on t_a. However, it is important to note that the analysis of each gridcolumn actually occurs at the mean observation time of the gridcolumn by a single 5 min duration MODIS granule. When multiple satellites/multiple orbits observe the same gridcolumn within the time window, as happens commonly at high latitudes, the MODIS granule that contributes observations closest to t_a is selected and so the time mismatch with the nominal analysis time is minimized. Even with this selection, the primary timing error is that the analysis is actually valid at a time displaced up to 90 min from the nominal analysis time. A secondary timing error is from interpolation of the GEOS-5 model background state to the mean observation time of the gridcolumn in order to form the prior. In time-averaged studies, we expect these two sources of timing error to act as additional forms of random error.

The cloud data consist of high spatial resolution retrievals of CO_2-slicing cloud-top pressure p_c, 10.8–11.3 μm brightness temperature T_b and visible cloud optical thickness (COT) τ from the the MODIS instrument aboard the Earth Observing Satellites *Terra* and *Aqua*. p_c and T_b are available at a nominal 5 km nadir pixel resolution, while τ is available at a higher 1 km resolution, all from the Collection 5.1 'MxD06_L2' Level 2 cloud granules (e.g. Yang *et al.*, 2007; Wind *et al.*, 2010). The basic quantum of analysis is the GCM-gridcolumn, so pixels within the analysis window are associated with the GCM-gridcolumn in which they fall. Clearly, the number of 1 km pixels within a GCM-gridcolumn can be large (≈ 625 for a tropical gridcolumn at $1/4^\circ$ model resolution). It is precisely the statistical cloud information contained in this collection of pixels that we wish to take advantage of in our CDA procedure.

The cloud-top pressure is only used if $p_c \leq 550$ hPa, since the CO_2-slicing algorithm performs poorly for clouds below this level. The cloud optical thickness is only used for 'daytime' gridcolumns, defined as those for which all contained pixels have a cosine of solar zenith angle ≥ 0.15. In fact, the MODIS cloud optical property retrievals, including cloud optical thickness, are not performed for cosines of solar zenith angle below this limit.

All 5 km data are copied to the underlying 1 km grid, so that all gridcolumn collection and pixel analysis is done at 1 km. We do this because we want to study the co-distribution of τ with p_c and T_b and because the plan for the soon-to-be-released Collection 6 is to produce p_c and T_b at 1 km.

(A) For 'daytime' pixels, the 1 km MODIS Cloud Mask Scientific Data Set (SDS) is used to classify pixels as either cloudy (those that are 'confidently or probably cloudy') or clear. For consistency, those pixels that are clear according to this measure have their p_c, T_b and τ set to zero. (This means that (i) copies of 5 km p_c and T_b at 1 km are only retained for 1 km cloudy pixels and (ii) T_b is the brightness temperature for cloudy pixels, not all pixels.)

After this clear pixel treatment, any pixels that have undefined values in τ or in both of p_c and T_b are discarded. These may include, for example, attempted but failed retrievals and so-called 'clear-sky restoral' pixels. Such anomalous retrievals are not easy to forward-model and so currently must be discarded.

In particular, the Collection 5.1 MODIS optical property algorithm eliminates some pixels in an 'edge-restoral' process (see Zhang and Platnick, 2011; Pincus *et al.*, 2012). These are pixels that MODIS detects as being on the edge of clouds and therefore discards from the optical processing on the basis of what is a fundamentally conservative retrieval philosophy, namely one that would rather provide a smaller number of well-defined and reasonably accurate retrievals than a larger number of cloud pixels containing potentially dubious data. This is a completely reasonable retrieval philosophy with distinct benefits for many users of the MODIS cloud optical properties, but for our particular purpose it does remove some pixels that could potentially have useful information content in our CDA application. The MODIS algorithm does flag such undefined 'edge-restoral' pixels, so this flag contains potentially useful information for us. We are not, however, currently able to make use of it, because the forward modelling of cloud edge effects requires a GCSM that includes horizontal spatial coherence. This is beyond the current capability of our algorithms, but see Wind *et al.* (2013) for an example of some progress in this direction.

(B) For 'night-time' pixels, the 1 km MODIS cloud mask SDS is either not available or not reliable and the 5 km cloud mask is only representative of the central 1 km pixel of the associated 5 × 5 box. Furthermore, for computational efficiency reasons, we currently do not read the associated MxD35_L2 cloud mask granule to obtain any further information on cloudiness at 1 km. Therefore, for night-time pixels, we currently resort to a cloudy mask if p_c is defined and a clear mask if it is undefined. Again, for consistency, the p_c and T_b of clear pixels are set to zero. The cloudy pixels have their p_c values reset to undefined if they are considered unreliable (for low clouds, as described earlier) and are discarded completely if their T_b is also undefined.

In summary, we end up with a collection of pixels for each model gridcolumn. Each pixel is either clear or cloudy. Clear pixels have zero p_c and T_b, while cloudy pixels must have a valid positive p_c or T_b (or both). Pixels in daytime gridcolumns also have a τ value: zero for clear pixels and positive for cloudy pixels. Finally, if the area covered by a gridcolumn's collection of pixels does not cover at least half of the footprint of the gridcolumn, then the whole gridcolumn is discarded. This weeds out gridcolumns that are undersampled by MODIS data.

2.5. *Forward modelling of observations*

Assume we are given an ensemble of generated subcolumns of a gridcolumn, as discussed in section 2.2. The goal is to produce a forward-modelled estimate of MODIS-retrieved p_c, T_b and τ for each subcolumn, so that the statistical properties of the ensemble of such subcolumn observables may be compared with the statistics of the MODIS observations associated with the gridcolumn.

The first step is to calculate the visible COT τ_k of each layer k of each subcolumn. (The details of the calculation are not particularly important. They broadly follow the GEOS-5 visible COT calculation, with cloud condensate amount per equation (2), a liquid/ice split as in section 2.8 and effective radii that are functions of pressure only.) The total subcolumn COT τ is just the sum of τ_k over all the layers of the subcolumn. If this sum is less than a nadir-adjusted detection threshold of $0.3\mu_{\mathrm{sat}}$ optical depths (where μ_{sat} is the cosine of the satellite zenith angle), then the subcolumn is considered 'MODIS-clear', since it falls below the nominal MODIS optical depth detection limit, and τ is reset to zero, together with p_c and T_b, for consistency with the treatment of the observations.

Next, for 'MODIS-cloudy' subcolumns, the CO_2-slicing cloud-top pressure p_c is calculated using the same simple approximation used by the Cloud Feedback Model Intercomparison Project (CFMIP) Observation Simulator Package (COSP) MODIS simulator (Bodas-Salcedo *et al.*, 2011). The details are given in Appendix C1. The brightness temperature, T_b, is only simulated if p_c is deemed unreliable, namely if the above calculated p_c exceeds 550 hPa, as discussed in section 2.4. In this case, p_c is set to undefined and the 10.5 μm T_b is evaluated with an IR-only version of the all-sky brightness temperature calculation used in the COSP International Satellite Cloud Climatology Project (ISCCP) Clouds and Radiances Using Subgrid Cloud Overlap Profile Sampler (SCOPS) (ICARUS) algorithm, as in the Appendix of Klein and Jakob (1999). Details are given in our Appendix C2.

After simulation of p_c, T_b and τ, as above, cloudy pixels have valid values clipped to $[1, 1100]$ hPa, $[150, 350]$ K and $[0, 100]$, respectively, to mirror the valid ranges produced by MODIS. These limits are only occasionally hit for the upper COT limit (see figure 2 in Part 2). In this sense, our forward model mirrors the behaviour of the retrieval.

As a final step, observation errors sampled from a prescribed observation-error PDF can be added to the simulated observations. The specification of observation errors is an important ingredient of any data-assimilation algorithm rooted in estimation theory. In our context, observation errors include state-dependent errors in the MODIS cloud retrievals, as well as errors arising from our forward operators. In a follow-on study to Wind *et al.* (2013), we are currently characterizing these observation errors in an Observing System Simulation Experiment (OSSE) context using a MODIS cloud retrieval simulator. For the present study, no observation-error term is added.

2.6. *Evaluation of the likelihood*

Given a parameter set $\boldsymbol{\alpha}$, we must evaluate the posteriori PDF in (1), or, in practice, its natural logarithm, $\ln p(\boldsymbol{\alpha}|\boldsymbol{y}) = \ln p(\boldsymbol{y}|\boldsymbol{\alpha}) + \ln p(\boldsymbol{\alpha})$, up to a constant independent of $\boldsymbol{\alpha}$. (The use of the logarithm here mirrors the traditional treatment of the minimization problem and leads to a more careful treatment of overflow issues for very large probabilities.) Firstly, if any parameter is outside acceptable bounds, as specified in section 2.3.2, then zero probability is returned. Secondly, the prior $p(\boldsymbol{\alpha})$ is evaluated per section 2.3.3. This leaves the likelihood, $p(\boldsymbol{y}|\boldsymbol{\alpha})$, which (loosely speaking) is the probability of observing $\boldsymbol{y}$ from a gridcolumn in state $\boldsymbol{\alpha}$. We will be more precise shortly.

Note that $\boldsymbol{y}$ above refers to the vector of all observations for the gridcolumn, comprising multiple pixels and multiple properties per pixel. Consider a gridcolumn with N pixels assigned to it by the observation-gathering routine (section 2.4). We assume that the observations at different pixels are independent and identically distributed (i.i.d.), so that

$$\mathcal{L} \equiv \ln p(\boldsymbol{y}|\boldsymbol{\alpha}) \approx \sum_{n=1}^{N} \ln \hat{p}(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha}), \tag{6}$$

where $\hat{\boldsymbol{y}}_n$ is the vector of properties observed for pixel n and $\hat{p}$ is the PDF applying to a single pixel (and common to all pixels).

In reality, we know that nearby pixels will not be independent, but our statistical model is currently not able to take account of horizontal coherence within a gridcolumn (the independent column approximation or ICA). In practice, this means that we apply more observations to the evaluation of $\ln p(\boldsymbol{y}|\boldsymbol{\alpha})$ than there are truly independent observations. This will give $p(\boldsymbol{y}|\boldsymbol{\alpha})$ an incorrect magnitude, but, ignoring the prior term $p(\boldsymbol{\alpha})$ for now, it should not seriously bias the estimation of $\boldsymbol{\alpha}$ if the variability within a gridcolumn is more or less homogeneous. The non-independence issue may be problematic, however, if the gridcolumn footprints are sufficiently large to regularly contain significant gradients in the scale of variability, in which case it will create a preferential bias between large and small scales of variability. We will not consider this possibility further, because the nominal 1/4° GCM model resolution used in our studies is small and because it is not yet computationally feasible to employ a GCSM with horizontal spatial coherence.

The presence of the prior term $p(\boldsymbol{\alpha})$ creates an additional problem. The posterior PDF $p(\boldsymbol{\alpha}|\boldsymbol{y})$ is the product of the likelihood and prior terms. Errors in the likelihood term therefore indirectly introduce effective errors in the proper *influence* of the prior information. In particular, if the gridcolumn is oversampled with observations compared with the horizontal correlation scale (the typical separation of independent observations) then the naive assumption of independence will have the effect of increasing the influence of the likelihood term erroneously compared with the prior, or effectively de-weighting prior information. (Using all the pixels, rather than just a subset representing independent observations, will cause the term $\sum_{n=1}^{N} \ln \hat{p}(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha})$ to exert an unrealistic influence compared with $\ln p(\boldsymbol{\alpha})$.) The influence of the prior information is therefore indirectly and erroneously dependent on the size of the pixel compared with the horizontal correlation scale. In practice, this error will be introduced in two ways: (i) because of the spatial and temporal variation of the horizontal correlation scale with the synoptic condition and (ii) because of the increase in pixel size towards the extrema of the scan lines. While some simple correction for (ii) might be possible, (i) will require an earnest treatment of horizontal spatial coherence, which is currently not computationally feasible and is deferred until future work. Thus, we proceed with our independent observations assumption, mindful that some geographical and scan-angle-dependent biases may be introduced.

Returning to (6), in our case, $\hat{\boldsymbol{y}}_n \equiv (\ln\tau, \omega)_n$, where ω is either p_{c} if $p_{\mathrm{c}} \leq 550\,\mathrm{hPa}$ or T_{b} otherwise. We prefer to use the logarithm of τ, rather than τ itself, because the former is more symmetrically distributed. The meaning of $\hat{p}(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha})$ is non-trivial, because we must account for both cloudy pixels ($\tau, \omega > 0$) and clear pixels ($\tau, \omega = 0$). The latter are produced not only by very thin clouds that fall below the MODIS detection limit, but in large part from truly clear portions of the gridcolumn. (In this sense, clear pixels are akin to a type of 'left-censored' observation.) The precise meaning of the likelihood $\hat{p}(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha})$ in (6) is explained in Appendix B1, namely that it is $P_\circ(\boldsymbol{\alpha})$ for clear pixels and $P_\bullet(\boldsymbol{\alpha})\,\hat{p}_\bullet(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha})$ for cloudy pixels, where $P_\circ(\boldsymbol{\alpha})$ is the likelihood of a clear pixel, $P_\bullet(\boldsymbol{\alpha}) = 1 - P_\circ(\boldsymbol{\alpha})$ is the likelihood of a cloudy pixel and $\hat{p}_\bullet(\hat{\boldsymbol{y}}|\boldsymbol{\alpha})$ is the likelihood density conditioned for cloudy pixels only, such that its integral over all cloudy $\hat{\boldsymbol{y}}$ is one. Note that $P_\circ(\boldsymbol{\alpha})$ and $P_\bullet(\boldsymbol{\alpha})$ are pure probabilities, while $\hat{p}_\bullet(\hat{\boldsymbol{y}}|\boldsymbol{\alpha})$ is a probability density with respect to $\hat{\boldsymbol{y}}$.

To proceed, we must know how to evaluate the likelihoods $P_\circ(\boldsymbol{\alpha})$ and $\hat{p}_\bullet(\hat{\boldsymbol{y}}|\boldsymbol{\alpha})$ at any test point $\boldsymbol{\alpha}$ in parameter space. Ideally, an analytic expression for these two quantities would be provided. However, for the complex multi-layer GCSM of section 2.2 and the complex observation operators for p_{c}, T_{b} and τ of section 2.5, it is effectively impossible to provide or evaluate such analytic expressions. More precisely, for $P_\circ(\boldsymbol{\alpha})$ and the Gaussian copula vertical correlation model described earlier, an analytic expression is available (see N08), but it is computationally too expensive to use it. The term $\hat{p}_\bullet(\hat{\boldsymbol{y}}|\boldsymbol{\alpha})$ is even worse, being expressible only in terms of a complex multidimensional integral that would need numerical evaluation anyway and would be likewise computationally prohibitive. Rather, following the example of N08, we perform a Monte Carlo evaluation of both $P_\circ(\boldsymbol{\alpha})$ and $\hat{p}_\bullet(\hat{\boldsymbol{y}}|\boldsymbol{\alpha})$ using an ensemble of N_{sim} subcolumns generated at $\boldsymbol{\alpha}$ (see section 2.2), as described below.

Say there are $N_\circ$ clear pixels and $N_\bullet \equiv N - N_\circ$ cloudy pixels in the observations. The clear pixels contribute

$$\mathcal{L}_\circ = N_\circ \ln P_\circ(\boldsymbol{\alpha}), \tag{7}$$

to (6). $P_\circ(\boldsymbol{\alpha})$ is estimated by the ratio of the number of clear simulated subcolumns to the total number of simulated subcolumns N_{sim} for subcolumn generation at $\boldsymbol{\alpha}$. Conversely, the cloudy pixels contribute

$$\mathcal{L}_\bullet = N_\bullet \ln P_\bullet(\boldsymbol{\alpha}) + \sum_{n\in\bullet} \ln \hat{p}_\bullet(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha}), \tag{8}$$

where '$n \in \bullet$' indexes cloudy pixels only. Before proceeding, there is a further level of conditioning that must be addressed. Namely, the cloudy pixels are decomposed further into two sets: those for which $\omega = p_{\mathrm{c}}$, indexed as '$n \in \bullet p_{\mathrm{c}}$', and the complement for which $\omega = T_{\mathrm{b}}$, indexed as '$n \in \bullet T_{\mathrm{b}}$'. Then, using the lines of the argumentation above, we can rewrite (8) as

$$\begin{aligned}\mathcal{L}_\bullet =& N_\bullet \ln P_\bullet(\boldsymbol{\alpha}) + N_{\bullet p_{\mathrm{c}}} \ln P(p_{\mathrm{c}}|\bullet, \boldsymbol{\alpha}) \\ &+ \sum_{n\in\bullet p_{\mathrm{c}}} \ln \hat{p}_{\bullet p_{\mathrm{c}}}((\ln\tau, p_{\mathrm{c}})_n|\boldsymbol{\alpha}) + N_{\bullet T_{\mathrm{b}}} \ln P(T_{\mathrm{b}}|\bullet, \boldsymbol{\alpha}) \\ &+ \sum_{n\in\bullet T_{\mathrm{b}}} \ln \hat{p}_{\bullet T_{\mathrm{b}}}((\ln\tau, T_{\mathrm{b}})_n|\boldsymbol{\alpha}),\end{aligned} \tag{9}$$

where $N_{\bullet p_\mathrm{c}}$ and $N_{\bullet T_\mathrm{b}}$ are the number of p_c-cloudy and T_b-cloudy pixels (with $N_{\bullet p_\mathrm{c}} + N_{\bullet T_\mathrm{b}} = N_\bullet$) and $P(p_\mathrm{c}|\bullet, \boldsymbol{\alpha})$ and $P(T_\mathrm{b}|\bullet, \boldsymbol{\alpha})$ are the likelihoods of such pixels, conditioned on cloudiness in general. In practice, these likelihoods are estimated as the ratios of the number of p_c-cloudy and T_b-cloudy generated subcolumns, respectively, to the total number of cloudy generated subcolumns at $\boldsymbol{\alpha}$ (section 2.5). The term $\hat{p}_{\bullet p_\mathrm{c}}((\ln\tau, p_\mathrm{c})|\boldsymbol{\alpha})$ is the likelihood density conditioned for p_c-cloudy pixels only, such that its integral over the domain on which both $\tau > 0$ and $p_\mathrm{c} > 0$ is one. In practice it is estimated from the collection of all subcolumns generated at $\boldsymbol{\alpha}$ that are both cloudy and have p_c available, as described in the next paragraph. Analogous comments apply for $\hat{p}_{\bullet T_\mathrm{b}}((\ln\tau, T_\mathrm{b})|\boldsymbol{\alpha})$.

The likelihood $\hat{p}_{\bullet p_\mathrm{c}}((\ln\tau, p_\mathrm{c})|\boldsymbol{\alpha})$ is estimated by the empirical PDF of the subset of subcolumns generated at $\boldsymbol{\alpha}$ that are both cloudy and have p_c available. The input is the set of $(\tau > 0, p_\mathrm{c} > 0)$ pairs produced by the subcolumn generation, which we will denote as $\{(\tau, p_\mathrm{c})\}$. The number of such pairs will be denoted as $N^{\mathrm{sim}}_{\bullet p_\mathrm{c}}$. Several possible PDF construction methods are used. The most basic is a single two-dimensional (2D) Gaussian distribution using the sample mean and covariance of $\{(\ln\tau, p_\mathrm{c})\}$. This has the advantage of simplicity and speed, but the disadvantage of being unimodal, symmetric and of fixed parametric (Gaussian) form. A much less restrictive method is to use a kernel density estimate (KDE) with Gaussian kernels. This KDE can represent complex, multimodal distributions and is our default method. A few details of the KDE application are given in Appendix B3. There is a further method that captures the simplicity and efficiency of the single Gaussian PDF, but relaxes the latter's Gaussian marginals for $\ln\tau$ and p_c. It is the Gaussian copula (GCOP) discussed in N08, which allows the coupling of arbitrary marginals for $\ln\tau$ and p_c. We have not yet tried this GCOP method, but it may prove to be useful and efficient, especially in cases where the $(\ln\tau, p_\mathrm{c})$ distribution is unimodal but non-Gaussian.

Note that all the above applies analogously for the T_b-cloudy likelihood $\hat{p}_{\bullet T_\mathrm{b}}((\ln\tau, T_\mathrm{b})|\boldsymbol{\alpha})$ in the third line of (9). We decided to use an 'either/or' approach for p_c and T_b, namely opting to use the brightness temperature only when $p_\mathrm{c} > 550$ hPa. There is nothing to stop the retrieval and simulation of brightness temperature all of the time. In that case, one would need to use a 3D $(\ln\tau, p_\mathrm{c}, T_\mathrm{b})$ PDF when p_c was available. A 3D KDE is somewhat more complicated and expensive, so we have decided to use the either/or approach for now. However, higher-dimensional observation spaces definitely warrant further investigation, particularly since they may be appropriate for multi-spectral cloudy radiance assimilation. Perhaps a higher-dimensional Gaussian or GCOP may prove acceptable for such cases. Nevertheless, it is likely that the usefulness of our method will be restricted to analyses with a small number of observables per pixel, be they retrieved quantities or radiances.

This completes the basic description of the likelihood evaluation. Additional technical details regarding the efficient computational implementation of the algorithm are given in Appendix B2.

2.7. Monte Carlo characterization of the posterior PDF

In the previous section, we presented our method for evaluating $p(\boldsymbol{\alpha}|\boldsymbol{y})$ at a particular $\boldsymbol{\alpha}$. We now outline the method used to find the set of parameters $\boldsymbol{\alpha}$ that maximizes this posterior PDF and also characterizes the PDF more fully, so that error estimates for the optimal $\boldsymbol{\alpha}$ may be provided or we may look for multiple modes, for example. The method is a form of MCMC. Such methods make quasi-random jumps around parameter space, such that, as the number of jumps becomes large, the collection of sampled $\boldsymbol{\alpha}$ is representative of the target PDF. The application of MCMC to Bayesian analysis is discussed thoroughly in Gelman *et al.* (2004). Posselt (2013) also provides a nice introduction, with applications to satellite retrieval and model parameter estimation. Several of the benefits of using this sort of Monte Carlo method were discussed in section 2.1.

2.7.1. Background: the Metropolis–Hastings algorithm

First, we will outline the traditional Metropolis–Hastings (MH) approach to MCMC. This is necessary to explain the 'curse of dimensionality' problem with the MH algorithm, which justifies our use of an alternative and less commonly known variant of MCMC, the multiple-try Metropolis (MTM) method of Liu *et al.* (2000), described later in this section. The following discussion of the MH algorithm serves as a good baseline and introduction to the MTM method described later.

The target distribution we wish to sample, the posteriori PDF $p(\boldsymbol{\alpha}|\boldsymbol{y})$, will be denoted simply as $\pi(\boldsymbol{\alpha})$. More precisely, $\pi(\boldsymbol{\alpha}) \equiv p(\boldsymbol{y}|\boldsymbol{\alpha})\, p(\boldsymbol{\alpha})$, the product of the prior and likelihood terms, since the other factor $p^{-1}(\boldsymbol{y})$ in Bayes' formula is independent of the parameters $\boldsymbol{\alpha}$ we are maximizing against. The MH algorithm produces a Markov chain $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n$, such that, as $n \to \infty$, the collection of points in the chain is an accurate statistical sample from $\pi(\boldsymbol{\alpha})$. (In practice, for finite chains, an optional initial portion of the chain, the 'burn-in' portion that remembers $\boldsymbol{\alpha}_0$, may be discarded. This burn-in period is also typically used to tune the first-guess proposal covariance, based on the initial sampling of the target distribution during the burn-in.) The procedure for going from $\boldsymbol{\alpha}_t$ to $\boldsymbol{\alpha}_{t+1}$ is as follows: produce a trial point $\boldsymbol{\alpha}'$ from a proposal PDF $q(\boldsymbol{\alpha}'; \boldsymbol{\alpha}_t)$. Calculate the 'MH ratio':

$$r_{\mathrm{MH}} = \min\left\{ \frac{\pi(\boldsymbol{\alpha}')\, q(\boldsymbol{\alpha}_t; \boldsymbol{\alpha}')}{\pi(\boldsymbol{\alpha}_t)\, q(\boldsymbol{\alpha}'; \boldsymbol{\alpha}_t)}, 1 \right\}. \tag{10}$$

This is the probability of acceptance of the jump from $\boldsymbol{\alpha}_t$ to $\boldsymbol{\alpha}'$. Thus, if a random number drawn uniformly from $[0, 1]$ is less than r_{MH}, then $\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}'$, otherwise the old point is retained, $\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t$. The proposal distribution q remains to be specified. If it is symmetric, then the q terms cancel. This is the case with a multivariate Gaussian distribution (which is also our choice):

$$q(\boldsymbol{\alpha}'; \boldsymbol{\alpha}_t) = \frac{\mathrm{e}^{-\frac{1}{2}(\boldsymbol{\alpha}'-\boldsymbol{\alpha}_t)^T \boldsymbol{\Sigma}_q^{-1} (\boldsymbol{\alpha}'-\boldsymbol{\alpha}_t)}}{\{(2\pi)^d |\boldsymbol{\Sigma}_q|\}^{1/2}}, \tag{11}$$

where $\boldsymbol{\Sigma}_q$ is the proposal covariance matrix and d is the number of dimensions. Thus, for the symmetric q case, (10) reduces to $r_{\mathrm{MH}} = \min\{\pi(\boldsymbol{\alpha}')/\pi(\boldsymbol{\alpha}_t), 1\}$. Clearly, if the trial jump moves to higher target probability then it is unconditionally accepted. If, however, $\pi(\boldsymbol{\alpha}')$ is reduced from $\pi(\boldsymbol{\alpha}_t)$, the jump is only conditionally accepted, with a probability proportional to $\pi(\boldsymbol{\alpha}')/\pi(\boldsymbol{\alpha}_t)$. Thus, while large reductions are unlikely, they are still possible, a fact that permits the algorithm to leave the vicinity of a local maximum and potentially to find the global maximum.

Note that advancing from $\boldsymbol{\alpha}_t$ to $\boldsymbol{\alpha}_{t+1}$ is a two-step procedure: first proposing a trial $\boldsymbol{\alpha}'$, then conditionally accepting it (or not). Thus, the probability of a chain transition from a point $\boldsymbol{\alpha}$ to a new (i.e. different) point in a small volume $\mathrm{d}\boldsymbol{\alpha}'$ containing $\boldsymbol{\alpha}'$, denoted $a(\boldsymbol{\alpha}'; \boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha}'$, is not the MH ratio for the proposed transition, but rather $q(\boldsymbol{\alpha}'; \boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha}'$ times the ratio. For symmetric q, then, $a(\boldsymbol{\alpha}'; \boldsymbol{\alpha}) = q(\boldsymbol{\alpha}'; \boldsymbol{\alpha}) \min\{\pi(\boldsymbol{\alpha}')/\pi(\boldsymbol{\alpha}), 1\}$ and so $\pi(\boldsymbol{\alpha})\, a(\boldsymbol{\alpha}'; \boldsymbol{\alpha}) = q(\boldsymbol{\alpha}'; \boldsymbol{\alpha}) \min\{\pi(\boldsymbol{\alpha}'), \pi(\boldsymbol{\alpha})\} = \pi(\boldsymbol{\alpha}')\, a(\boldsymbol{\alpha}; \boldsymbol{\alpha}')$, or, written in another way,

$$\frac{a(\boldsymbol{\alpha}'; \boldsymbol{\alpha})}{a(\boldsymbol{\alpha}; \boldsymbol{\alpha}')} = \frac{\pi(\boldsymbol{\alpha}')}{\pi(\boldsymbol{\alpha})}. \tag{12}$$

This means, for example, that if $\pi(\boldsymbol{\alpha}')$ is twice $\pi(\boldsymbol{\alpha})$ then the probability of moving from $\boldsymbol{\alpha}$ to $\boldsymbol{\alpha}'$ is twice that of moving in the reverse direction. This symmetry is known as 'detailed balance' and ensures that the Markov chain converges to sampling $\pi(\boldsymbol{\alpha})$ accurately for a long enough chain (e.g. Andrieu *et al.*, 2003).

Now, imagine that $\boldsymbol{\alpha}_t$ is in the vicinity of either a local maximum or the global maximum of $\pi(\boldsymbol{\alpha})$. In the former case, the goal is to jump off the maximum and search for the global maximum. In either case, however, as discussed in section 2.1, the goal is also to sample the entire PDF and not just find the global maximum. First, say that the typical size of a trial jump $|\boldsymbol{\alpha}' - \boldsymbol{\alpha}_t|$ is very small compared with the underlying local scale of $\pi(\boldsymbol{\alpha})$. Then individual jumps will produce very small changes in $\pi(\boldsymbol{\alpha})$ and so the acceptance rate will be very high. (The 'acceptance rate' is the fraction of accepted trials in some finite sequence of the chain.) This will lead to a slow, near-random walk, with small steps, through the $\boldsymbol{\alpha}$ domain and a very inefficient sampling of $\pi(\boldsymbol{\alpha})$. Conversely, say that the typical size of a trial jump is large relative to the underlying local scale of $\pi(\boldsymbol{\alpha})$. In this case, if we are in the vicinity of a local maximum or the global maximum, a typical trial jump will severely reduce $\pi(\boldsymbol{\alpha}')$ from $\pi(\boldsymbol{\alpha}_t)$ and so the acceptance rate will be very small. This will again lead to very inefficient sampling of $\pi(\boldsymbol{\alpha})$ (and also potentially a very inefficient search for the global maximum if we are on a local maximum). Based on this argument, there ought to be some optimal intermediate acceptance rate, associated with some intermediate and optimal typical trial jump size, that leads to the most efficient sampling of $\pi(\boldsymbol{\alpha})$.

According to a number of studies (see Posselt, 2013, for references), for large d the optimal acceptance rate is about 20% (and 23.4%, in particular, for a target $\pi(\boldsymbol{\alpha})$ with i.i.d. components of $\boldsymbol{\alpha}$). According to Roberts and Rosenthal (2001), the optimal $\boldsymbol{\Sigma}_q$ is proportional to the covariance of the target distribution, $\boldsymbol{\Sigma}_\pi$, although the latter is not generally known *a priori*. For a multivariate Gaussian $\pi(\boldsymbol{\alpha})$, the optimal $\boldsymbol{\Sigma}_q$ is

$$\boldsymbol{\Sigma}_q \approx \frac{(2.4)^2}{d} \boldsymbol{\Sigma}_\pi . \tag{13}$$

The reason for this effective scaling of the proposal jumps by $d^{-1/2}$ with respect to the scale of the target distribution is instructive for our discussion. Consider a multivariate Gaussian proposal distribution with a simple diagonal covariance $\boldsymbol{\Sigma}_q = \sigma_q^2 \mathbf{I}$. Then the proposal PDF is

$$q(\boldsymbol{\alpha}'; \boldsymbol{\alpha}_t) = \frac{\mathrm{e}^{-\frac{1}{2}|\boldsymbol{\alpha}'-\boldsymbol{\alpha}_t|^2/\sigma_q^2}}{\left(\sigma_q\sqrt{2\pi}\right)^d}, \tag{14}$$

and the margin of each α_i' is Gaussian with mean α_{ti} and standard deviation σ_q. Thus the scaled radial jump distance,

$$R \equiv \frac{|\boldsymbol{\alpha}' - \boldsymbol{\alpha}_t|}{\sigma_q} = \sqrt{\sum_{i=1}^{d} \left(\frac{\alpha_i' - \alpha_{ti}}{\sigma_q}\right)^2}, \tag{15}$$

is the root sum square of a set of independent standard Gaussians and therefore has a $\mathcal{X}(d)$ distribution, with mode $\sqrt{d-1}$. Thus the modal scaled jump distance varies as $\sqrt{d}$ for large d. Now, say the target distribution is centred on $\boldsymbol{\alpha}_t$, with covariance $\boldsymbol{\Sigma}_\pi = \sigma_\pi^2 \mathbf{I}$. It will therefore have the same form as (14), but with σ_π replacing σ_q, and so the modal proposal jump above will lead to an MH ratio $r_{\mathrm{MH}} = \pi(\boldsymbol{\alpha}')/\pi(\boldsymbol{\alpha}_t) = \exp(-\frac{1}{2}\sigma_q^2/\sigma_\pi^2 \times d)$. This, then, is why σ_q must be scaled by $d^{-1/2}$ with respect to σ_π, otherwise a typical proposal jump will give an extremely low MH ratio and acceptance rate, especially as the dimensionality of the problem becomes large.

Now the so-called 'curse of dimensionality' becomes clear: by using the scaling (13) to keep a reasonable acceptance rate, the scale of the proposal jumps, $\sigma_q \approx 2.4\,\sigma_\pi/\sqrt{d}$, becomes very small with respect to the scale of the target distribution σ_π as the dimensionality becomes large (or, in other words, each dimension (parameter) α_i becomes extremely oversampled). This is the reason that the MH algorithm becomes very slow for large-dimensional problems and is the background and motivation for the MTM method of Liu *et al.* (2000).

2.7.2. *The multiple-try Metropolis algorithm*

The MTM algorithm of Liu *et al.* (2000) seeks to mitigate the 'curse of dimensionality' of the MH algorithm by allowing larger trial jumps but still retaining a reasonable acceptance rate. Alternatively, we could say that by sampling the multi-dimensional parameter space more efficiently at each point in the Markov chain, faster convergence is achieved, i.e. fewer chain elements are required.

We use a simplified version of MTM that applies for a symmetric proposal $q(\boldsymbol{\alpha}'; \boldsymbol{\alpha}_t)$, such as the multivariate Gaussian (11) we are using. The algorithm is as follows.

- Make not one, but M independent trials $\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_M'$ from the proposal distribution $q(\,\cdot\,; \boldsymbol{\alpha}_t)$.
- Randomly select one of these, denoted as $\boldsymbol{\alpha}'$, from among the trials, but with a probability proportional to the $\pi(\,\cdot\,)$ of the points.
- Then from this $\boldsymbol{\alpha}'$ draw a further $M-1$ 'reference' trials $\boldsymbol{\alpha}_1^*, \ldots, \boldsymbol{\alpha}_{M-1}^*$ from $q(\,\cdot\,; \boldsymbol{\alpha}')$ and set an Mth reference point equal to the starting point, $\boldsymbol{\alpha}_M^* = \boldsymbol{\alpha}_t$.
- Finally, accept $\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}'$ with probability

$$r_{\mathrm{MTM}} = \min\left\{\frac{\pi(\boldsymbol{\alpha}_1') + \ldots + \pi(\boldsymbol{\alpha}_M')}{\pi(\boldsymbol{\alpha}_1^*) + \ldots + \pi(\boldsymbol{\alpha}_M^*)}, 1\right\}, \tag{16}$$

otherwise retain $\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t$.

Though more complicated, detailed balance (12) can also be proved for MTM.

The fact that a larger modal trial jump distance is possible, compared with MH, is not so easy to prove. By experimentation in the current CDA context (as described here in Part 1 and tested in Part 2), we have found that the MTM proposal covariance matrix $\boldsymbol{\Sigma}_q$ may be a factor $C = 32$ times the optimal MH value of (13). Details of this experimentation are found in Part 2. For the estimated covariance matrix $\boldsymbol{\Sigma}_\pi$ of (13), we use the following properties: for $\bar{S}$, we use the vertically correlated form in (5); for P_{L} and β, we also use the form (5) but with $\sigma_{P_{\mathrm{L}}}^2$ and σ_β^2 replacing $\sigma_{\bar{S}}^2$. Nominally, $\sigma_{P_{\mathrm{L}}} = 0.1$ and $\sigma_\beta = 0.1$, given that the valid ranges of P_{L} and β are [0, 1]. Also, as for the prior (section 2.3.3), the covariance between the $\bar{S}$, P_{L} and β parameters is taken as zero. This is a default assumption that will be changed once a cycling CDA with ensemble $\boldsymbol{\alpha}$ evolution is implemented.

The MTM implementation is performed carefully in terms of the logarithm of $\pi(\boldsymbol{\alpha})$, in order to avoid overflow problems. The length of the chain n and the number of trials M per point $\boldsymbol{\alpha}_t$ will be investigated as part of a series of sensitivity tests in Part 2. As a nominal guide, however, working reasonably well for the current context, we use $n = 200$ and $M = f_M \times M_*$, where $f_M = 1/2$ and $M_* = 3 \times (1000 - p_{\underline{\mathrm{lim}}})/L$, i.e. the number of parameters per layer, three (namely, $\bar{S}$, P_{L} and β), times the approximate number of effectively independent layers in the gridcolumn. See Part 2 for further discussion of these choices.

Although we mentioned above an optional burn-in period in which the specified first-guess proposal covariance is tuned, in our implementation we do not use such a burn-in period, but simply sample the target distribution using a fixed $n = 200$ point chain (where, as above, each point in the chain uses many ($M = 14$ in our case) trials from the proposal distribution, not one as in MH). We did experiment with various burn-in periods and proposal tuning cycles, but found that they did not improve the results and took more time.

Moreover, our fixed $n = 200$ point MTM chain employs no MCMC convergence criterion for the following reasons.

(1) Achieving convergence for MCMC is currently more of an art than a science. Theoretical convergence rates are currently few and far between and of limited practical use. Diagnostic methods offer some help, but none is conclusive in saying when the unknown target distribution

has been sampled satisfactorily. Furthermore, many of the diagnostic methods are expensive in themselves. Please see, for example, Cowles and Carlin (1996).

(2) We are not solving a single Bayesian inference problem as thoroughly as we possibly can, but trying to estimate a posterior mode and its error characteristics, rather approximately, for hundreds of thousands of problems (gridcolumns) at each analysis time. We therefore seek a very efficient and simple method of terminating the MCMC chain. Computational efficiency is paramount if one is seeking an algorithm for operational implementation.

2.8. *Other implementation details*

The CDA method described in this article has a very simple treatment of thermodynamic phase. Separate liquid water and ice contents, if present in the background state, are combined to a total condensate content. When the phase must be taken into account explicitly for a forward model calculation, such as for cloud optical thickness and infrared emissivity, the condensate is taken as all ice below $-35\,^{\circ}$C, all liquid above $0\,^{\circ}$C and a mixed fraction, linear in temperature, between these limits. This same weighting is also used to calculate the saturation vapour content $q_{\rm s}$ from the values over pure ice and liquid water and the maximum total saturation ratio $S_{\max}$ from values of 1.4 for ice and 1.1 for water. We regard these latter values as reasonable upper limits to the total saturation ratio, corresponding, respectively, to maximum liquid cloud water and cloud ice contents of 10% and 40% of the local saturation vapour content. The simplistic treatment of phase above is certainly a weak point of the current method that deserves further attention. The real phase split within mixed-phase clouds depends on many details of the local dynamic and thermodynamic environment (see, e.g., Noh *et al.*, 2013).

The GEOS-5 GCM used for the background state in this study has a simple latitude–longitude grid. Because this native grid would cause sampling problems near the poles, where the area of the gridcolumn footprint becomes very small, a more equal-area 'reduced longitude grid' was used for the CDA analysis described here and in Part 2. This grid retains the native I_M GCM longitudes per latitude for latitudes in the range [30°S, 30°N], but outside this range uses a reduced number, approximately $1+(I_M-1)\cos(\lambda)/\cos(30^{\circ})$, of longitudes per latitude, where λ is the latitude. The term 'gridcolumn' in Parts 1 and 2 actually refers to this reduced longitude grid. Aggregation and interpolation to this reduced grid are performed as necessary. Once GEOS-5 transitions to the close-to-equal-area cubed-sphere grid planned for GEOS-6, this reduced longitude grid will no longer be necessary.

3. A simple synthetic illustration of our method

In section 2, we provided a detailed description of our Bayesian MCMC CDA algorithm for assimilation of MODIS cloud data into realistic GCM-gridcolumns. A detailed testing of this algorithm with real MODIS data is provided in Part 2 of this series. However, in preparation, we conduct here a simpler and confidence-building test of the basic capability of the Monte Carlo algorithm in a more limited yet instructive context.

We will illustrate the ability of the algorithm to reconstruct realistic skewed triangle PDFs using censored synthetic cloud observations simulated from them. For the purpose of a straightforward illustration, we will limit our focus to the case of a single model layer, rather than a coupled vertical column of them. The model state is therefore a single triangular PDF characterized by the three parameters $(P_{\rm L},\beta,\bar{S})$.

Our illustration uses five case-study triangular PDFs in S (total saturation ratio): a clear case, three partially cloudy cases, with low, medium and high cloud fraction, and an overcast case. For each case, we take $N=625$ random samples from the distribution, approximately the number of 1 km pixels in a 1/4° gridcolumn. For each sample, the condensate-like observable $S_{\rm c}\equiv\max(S-1,0)$ is evaluated. (This observable is a simple analogue of τ for a single model layer.) The PDF from which these samples are drawn is called the 'truth' PDF, since it is used to specify synthetic $S_{\rm c}$ observations.

Each MCMC chain also *begins* from one of the five case-study triangles, but this one is called the 'background' PDF or initial condition. We thus perform 25 MCMC chains, one for every combination of background and truth PDFs from the five case-study triangles. Actually, we perform ten times as many chains, or 250 chains, since we make ten different realizations of the N synthetic observables for each truth triangle.

The most probable element from each MCMC chain is called the 'analysis' PDF and is compared with the truth PDF to determine the effectiveness of the Bayesian MCMC algorithm in recovering the truth using synthetic observations generated from it. It is important to remember, however, that the actual 'analysis' contains far more information than just the most probable PDF. Namely, the full *a posteriori* distribution of PDF parameters is obtained. We will see this in the subsequent result plots.

We have conducted this study with and without a prior in the parameters. However, the results we present below are *without* a prior, since this provides a more exacting test of the ability of the algorithm to retrieve the 'truth' from a potentially distant initial condition. Therefore, with no prior, the only non-observation constraint provided for S is that the triangle must lie wholly within $[0,S_{\max}]$, with a fixed $S_{\max}=1.1$. Similarly, the skewness parameter $P_{\rm L}$ is always constrained to $[0.1,0.9]$ and the width scaling parameter β to $[0,1]$.

For the complex multi-layer GCSM of section 2.2 and the complex observation operators for $p_{\rm c}$, $T_{\rm b}$ and τ of section 2.5, it is necessary to employ a KDE-based likelihood using simulated cloudy observations (section 2.6), since an analytic PDF is intractable in that case. However, in the present simpler illustrative study (a single triangle PDF and the simple observable $S_{\rm c}$), an analytic likelihood is easily derivable and is therefore used. Full details of this analytic likelihood are given in Appendix D. Using an analytic likelihood is preferable, since it provides a more direct test of the MCMC algorithm, without the added error due to a KDE approximation. We will address the choice of an analytic over KDE-based likelihood further in section 3.1 below.

Consequently, since we use an analytic PDF for $S_{\rm c}$, it is unnecessary to generate $N_{\rm sim}$ observables at each test point $\boldsymbol{\alpha}$ in parameter space (in order to form an empirical KDE-based likelihood of the observables.) Rather, the likelihood can be evaluated directly from the observations using the analytic PDF of $S_{\rm c}$ at $\boldsymbol{\alpha}$. See Appendix D for further details.

Several of the algorithmic parameters were changed because of the reduced dimensionality of this illustration compared with the full gridcolumn algorithm.

(1) We use $M=3$ trials per chain element, equal to the number of parameters.

(2) As above, we no longer need $N_{\rm sim}$. For $N_{\bullet\max}$ (see Appendix B2), we use a value in excess of N, the number of observations, meaning that every observation is used in evaluating the likelihood, not a random subset of them. There is no need to use a random subset for this illustration, since execution time is not at a premium and, again, removing unnecessary approximations focuses the study on the target: the capabilities of the MCMC algorithm.

(3) As per section 2.7.2, we use assumed target standard deviations $\sigma_{P_{\rm L}}=\sigma_{\beta}=\sigma_{\bar{S}}=0.1$. For the *full* gridcolumn case, based on (13), this translates to a proposal standard deviation in each dimension of $\sigma_q=\sqrt{32}\times 2.4/\sqrt{30\times 3}\times 0.1\approx 0.14$, where we have used the suggested proposal covariance amplification factor of $C=32$ (section 2.7.2) and where there are typically about 30 model layers below the tropopause (see section 2.3.2) and each has three triangle PDF parameters. However, for the current single-layer test, we will use $C=0.01$,

which translates to $\sigma_q = \sqrt{0.01} \times 2.4/\sqrt{3} \times 0.1 \approx 0.014$, about ten times smaller than the full gridcolumn case. We use this finer sampling of parameter space for the one-layer illustration, because we do not have the operational constraints of the full CDA algorithm and because we want to examine the posterior PDF in fine detail.

(4) However, as a result of this finer sampling of the posterior, we use a much larger number of chain elements $n = 10\,000$ than in the full gridcolumn $n = 200$ case. Since the number M of trial samples per chain element has been reduced from 14 (section 2.7.2) to 3 (item (1) above), this translates to an increase in posterior samplings by a factor of $10\,000 \times 3 \,/\, 200 \times 14$, or about 10 times. We will comment on the sensitivity to these parameter choices in section 3.1 below.

Of the 25 initial-condition/truth pairings, four diverse cases were selected for presentation here. The first of the ten realizations of these cases is illustrated in Figure 2. Case (a) has an initial condition starting from the truth and yields a very similar analysis; case (b) has excessive cloud fraction in the initial condition, but the analysis yields a near-truth PDF; case (c) starts from an unrealistic completely clear state but manages to restore the observed cloudiness, albeit with a less accurate analyzed PDF; and case (d) is a challenging case where the true cloud fraction is very small, so the synthetic observations provide more limited information. Nevertheless, the analysis from an unrealistic overcast initial condition yields a fairly reasonable analysis.

To investigate these cases in more detail, we look at details of the posterior distribution and MCMC chain behaviour in Figure 3. Because the parameter space is three-dimensional, we show the two-dimensional marginals for each of the parameter pairs: (P_L, β), $(P_\mathrm{L}, \bar{S})$ and $(\beta, \bar{S})$. For each marginal, the underlying colour plot is a kernel density estimate of the posterior marginal formed from all the chain elements. The blue cross shows the initial condition, the orange square the 'truth' and the large red dot the 'analysis', or, more precisely, the chain element with maximum *a posteriori* probability (MAP).

Consider first case (a), in the first column of Figure 3, for which the starting point of the chain is coincident with the truth. It might be expected in this case that the MAP is at the initial condition. In fact, the analysis is displaced slightly from the truth/initial condition, though certainly near the centre of the KDE of marginal posterior probability. This displacement occurs because of the finite sampling ($N = 625$ points) of the observable, so that the truth is only approximately represented by the synthetic observations. Nevertheless, noting the scale of the axes, the truth and analysis are indeed very close. Furthermore, as noted in the Introduction, since we have not only the MAP 'analysis' point (large red dot) but the full three-dimensional MCMC chain (from which a detailed KDE of the posterior distribution can be evaluated), we automatically have significant information to quantify the likely errors in the MAP 'analysis' parameters. The dependence of these errors on the true cloud fraction and on the number of observations (samples) is studied in section 3.1 below.

The other point that needs clarifying is that the analysis (large red dot) does not coincide with the maximum of the marginal KDE of the chain points (the underlying colour plot). This is because the KDE plot is of the *marginal*, i.e. it is the two-dimensional KDE of the chain points for the two parameters of the axes shown. This is an approximation to the full posterior probability integrated in the excluded third dimension. In other words, the maximum of the posterior probability in three dimensions does not coincide with its maximum after integration in one of the dimensions.

By contrast, for cases (b)−(d), *the initial point in the chain is a significant distance from the truth, but the marginal plots show that the chain is able to make its way to the region of significant posterior probability near the truth and and explore it.* As noted in the Introduction, *this is a significant advantage of the algorithm, being able to move out of regions of even zero posterior probability.* For example, the initial condition in case (c) is completely clear and therefore has *zero* probability of producing the observed cloudiness. Not only this, but the likelihood has all zero partial derivatives with respect to the parameters at this subsaturated initial condition. Nevertheless, the MCMC algorithm, with its non-gradient approach, is able to advance to a reasonable solution.

Despite this success, we note that the marginal KDEs for cases (b)−(d) show a rather elongated structure and that in case (c) a large error between truth and analysis occurs in P_L. To explain this, we first note that cases (b)−(d) all have cloud *only in the falling upper leg* of the truth triangle (see Figure 2). Thus, the probability density in the cloudy region, which determines the probability of occurrence of the cloudy observations ($S_\mathrm{c} > 0$), has a simple linear variation with S and is therefore fully specified by only two rather than the full three parameters of the triangular PDF. In particular, we need only specify the upper intersect S_H and the slope, which is just the negative of the height of the triangle $2/\Delta$ divided by the upper leg width $P_\mathrm{H}\Delta$, yielding $-2/\theta_\mathrm{H}$, where $\theta_\mathrm{H} \equiv P_\mathrm{H}\Delta^2$. Likewise, the clear observations, which all collapse to the censored $S_\mathrm{c} = 0$ observable, occur with a probability of $1 - f$ per observation, where f is the cloud fraction. Clearly, however, if the probability density $p_S(s)$ in the falling leg depends only on S_H and θ_H, then so does f, which is the integral of $p_S(s)$ from 1 to S_H. For these reasons, for the case of cloud only in the falling upper leg of the triangle, we might expect that any solution with the same S_H and product $P_\mathrm{H}\Delta^2$ will yield the same posterior probability.

While this is simply a heuristic explanation, it is fully validated by the detailed mathematical analysis presented in Appendix D. Namely, it is shown that the MAP solution for *infinite* observations drawn from a truth PDF with cloud only in the falling upper leg is *any* triangle with the same upper bound S_H and the same $\theta_\mathrm{H} \equiv P_\mathrm{H}\Delta^2$ as the truth, so long as its upper falling leg also contains the cloud fully. This is a classic identifiability problem. There is not just one theoretical MAP solution, but a whole family, having the same S_H and the same product $P_\mathrm{H}\Delta^2$ as the truth. This family of MAP solutions is shown as the magenta line in Figure 3 for cases (b)−(d). The dashed magenta lines are the limits of applicability in P_L of this theoretical solution, as explained in Appendix D.

Clearly the centre of the marginal KDE follows the magenta theoretical MAP solution quite closely for cases (b) and (c). Also, the MCMC analysis (large red dot) in case (c), while poor compared with the truth in P_L, is explained well by the identifiability issue, falling on the theoretical MAP curve. For case (d), the theoretical MAP curve is displaced somewhat from the marginal KDE, but remember that the theoretical curve is for perfect sampling (infinite observations), whereas we use only 625 observations. Clearly the cloudy observations ($S_\mathrm{c} > 0$) contain more information than the censored clear observations ($S_\mathrm{c} = 0$) and for case (d), with its very small cloud fraction, there are very few cloudy observations to constrain the solution. Therefore, it is not surprising that this case (d) has the marginal KDE that deviates most from the perfect sampling theory (see also section 3.1).

Figures 2 and 3 show only the first of ten separate realizations of observations from the truth. The other nine (not shown) confirm all of the conclusions above and, in particular, show variability of the analysis across the breadth of the perfect sampling curve for cases (b)−(d).

This completes our illustrative study of the Bayesian MCMC approach for a single-layer triangular PDF with a condensate-like censored observable. On the one hand, the illustration demonstrates the basic success of the MCMC algorithm in finding the theoretical MAP analysis, even from quite distant and very low probability initial conditions. On the other hand, the illustration has identified an identifiability problem for certain truth triangles. The cause of this problem relates to the simple piecewise linearity of the triangular PDF model, thereby sometimes reducing the

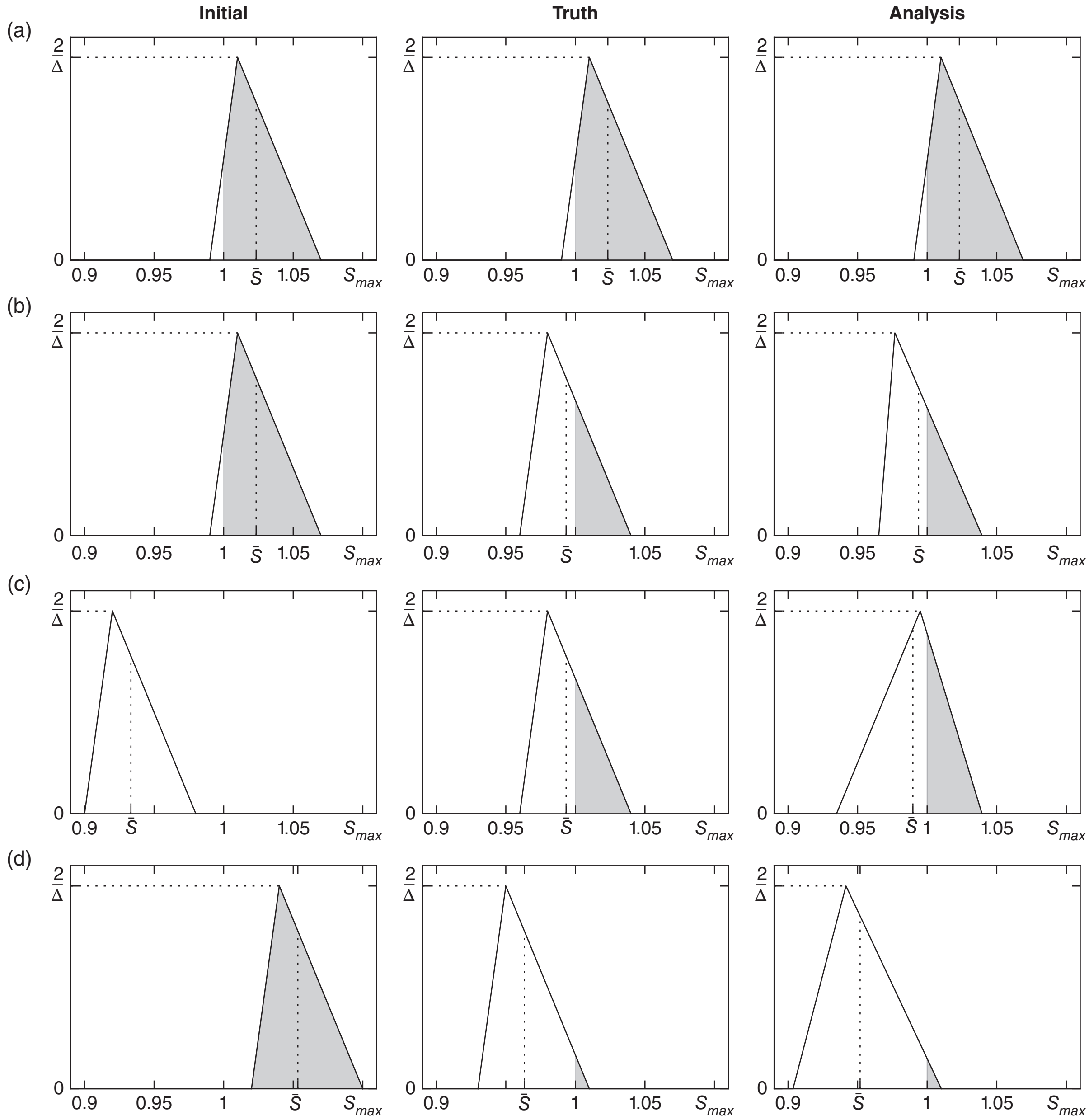

Figure 2. Four test cases for reconstruction of a triangular PDF from synthetic data (see text): 'initial' is the PDF from which the MCMC chain begins; 'truth' is the PDF from which the synthetic S_c observations were generated; 'analysis' is the PDF from the most probable element of the MCMC chain. Each PDF is in total saturation ratio S, with the mean shown as a vertical dotted line and the saturated portion ($S > 1$) shaded. Each PDF has a nominal vertical scaling – in reality, every PDF has a unit integral. Case (a) starts from the truth and yields a very similar analysis; (b) has excessive cloud fraction in the initial condition, but the analysis yields a near-truth PDF; (c) starts from an unrealistic completely clear state but manages to restore the observed cloudiness, albeit with a less accurate analyzed PDF; (d) is a challenging case where the true cloud fraction is very small, so the synthetic observations provide more limited information. Nevertheless, the analysis from an unrealistic overcast starting point yields a fairly reasonable analysis.

MAP solution to only two parameters, rather than the full three independent parameters. In retrospect, it might have been better to have used a smooth three-parameter PDF, such as the GEV distribution used in Norris *et al.* (2008). Then again, such more complex PDFs become analytically intractable very quickly.

In reality, this identifiability problem will be quickly mitigated in the multi-layer problem that is the main focus of Parts 1 and 2 of this series. This is because the sum over multiple layer triangles with any degree of vertical decorrelation quickly becomes smooth. For this reason, we do not expect this identifiability problem to manifest itself strongly in the results of Part 2 and there is no evidence that it does.

A more likely limitation for the method is the increased error for small cloud fractions, due to the limited constraining information then available. This is not a limitation of this method alone and is one reason why a prior PDF is required. Nevertheless, it is surprising how well the MAP solution performs for our low cloud-fraction case (d), even without a prior and even from a relatively distant initial condition (see also section 3.1).

Finally, it should be noted that the identifiability problems discussed above arise due to the censored observation operator S_c, which collapses to zero for all clear observations. If we were to add to the problem some ability to quantify the observed water vapour from the clear pixels, this would greatly enhance the information content available to constrain the analysis, especially in the single-layer case shown in this illustration. Indeed, we have conducted additional studies (not shown) replacing the censored observable S_c with the full moisture observable S. This causes the identifiability problem to disappear and the analysis is excellent. In this case, it is largely only the limited number of available observations that limits the accuracy of the analysis. This argues for the strong value-added potential of vapour observations, even if they are less spatially resolved.

3.1. Assumption/sensitivity studies

We conducted several additional studies, as a follow-up to the illustrative study above, to investigate the sensitivity of the Monte

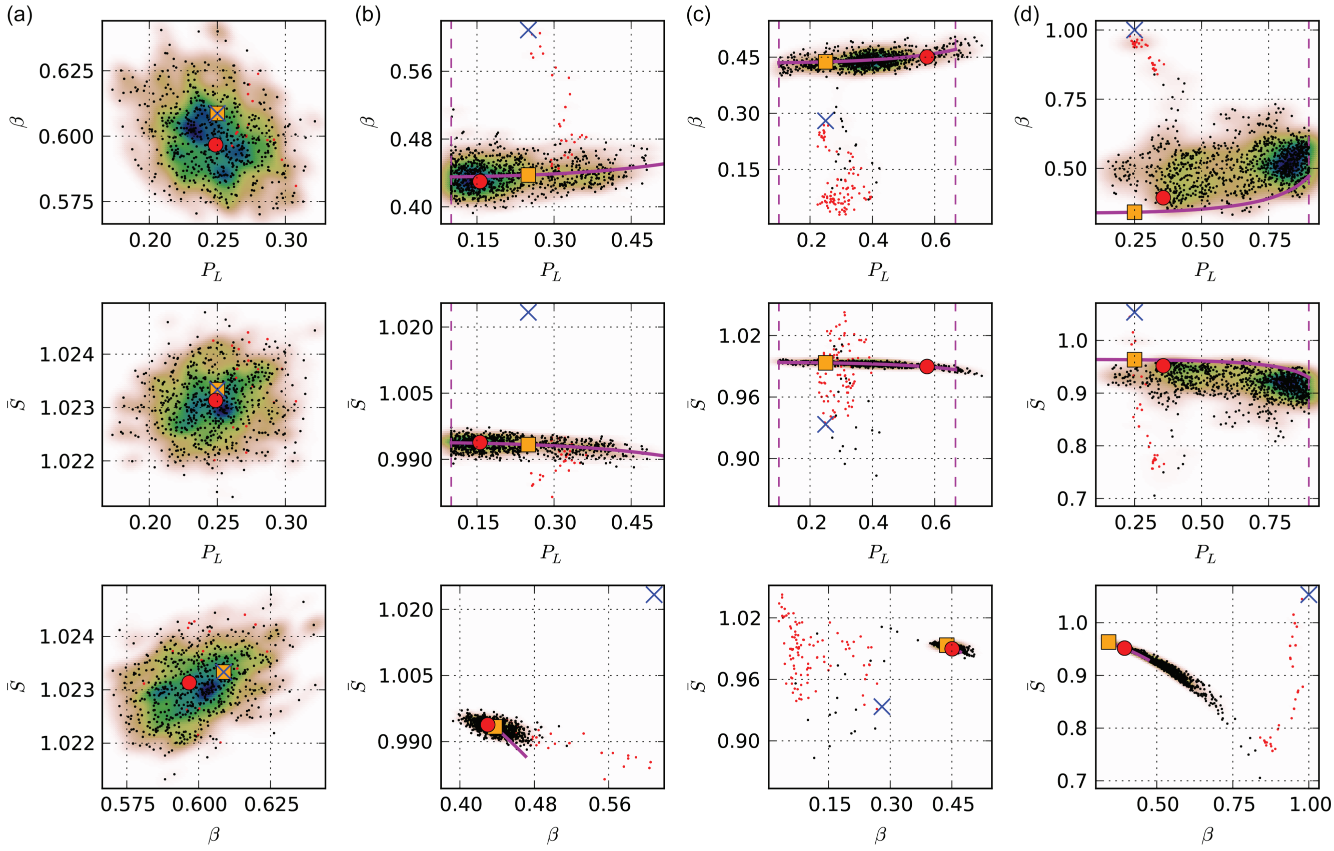

Figure 3. The plots show the two-dimensional marginals of the posterior PDF in $(P_L, \beta, \bar{S})$ for the cases (a)–(d) of Figure 2. The first, second and third rows show the marginals in (P_L, β), $(P_L, \bar{S})$ and $(\beta, \bar{S})$, respectively. Each underlying colour plot is a kernel density estimate of the marginal. The blue cross is the initial condition, the orange square is the 'truth' and the large red dot is the 'analysis'. The small red dots show the first 100 elements of the MCMC chain, while the black dots show every tenth element thereafter. The magenta solid and dashed lines for cases (b)–(d) show the *theoretical* maximum *a posteriori* solution and its limits of applicability (see text). Overall, the plots show both the strength of the MCMC algorithm in finding distant maxima and the errors in the analysis due to the limited information content of the synthetic observations and other algorithmic constraints, as described in the text. (Note the very different axis limits for each case.)

Carlo MAP analysis to various assumptions made, such as the use of an analytic versus KDE-based likelihood or various parameter choices, such as the number of observations N. The control will be the analytic likelihood and parameters noted in section 3, but we will use nine truth triangles spanning clear to overcast, instead of the earlier five, to sample the cloud fraction dependence better.

Figure 4 shows a comparison of the biases of the MAP values of five parameters ($\bar{S}$, Δ, P_L, S_H and θ_H) for the exact analytic triangular likelihood discussed in section 3 and the approximate KDE-based likelihood used by the full CDA algorithm (section 2.6).

We make the following observations.

(1) The truth triangles with modes $S_*^t \leq 1$ have large biases in P_L and Δ, since these triangles suffer from the identifiability problem discussed above. The biases of the parameters S_H and θ_H not affected by this identifiability issue are more acceptable.

(2) Large cloud fractions generally have smaller biases. This shows the value of near-full sampling of the truth triangles by cloudy observations and suggests the potential benefit of adding vapour observations in partially cloudy cases.

(3) In general, the approximate KDE-based likelihood solution approaches the exact analytic solution for S_H and θ_H as we increase the number of simulated points N_{sim} on which the KDE is based. The figure shows the improvement from $N_{\mathrm{sim}} = 64$ to 128 and the 256 case (not shown) is even better. A larger N_{sim} provides more cloudy samples with which to approximate the piecewise linear analytic likelihood with a KDE. Despite this reasonable approach of the KDE to analytic solution, we do note the somewhat different bias signatures between the two. This is most obvious for P_L, where the KDE solution tends to have larger positive biases for mid-range cloud fractions and larger negative biases for the overcast case. These differences are to be expected. There are fundamental limits to the use of a smooth KDE to model a piecewise linear triangular likelihood.

Figure 5 shows the dependence on N (the number of synthetic observations) of the median and interquartile range (IQR) bias, as a function of true cloud fraction, for the two parameters S_H and θ_H not affected by identifiability issues. These results are for the analytic likelihood and parameter values described in section 3. Each median/IQR is for the 90 cases (ten observable realizations for each truth triangle and nine different initial triangles) also used for Figure 4. The study is performed for three different N, namely 25, 625 and 2500, corresponding, respectively, to approximate numbers of 5 km, 1 km and 500 m pixels in a 1/4° gridbox. Clearly, from the figure, increasing the number of observations reduces both the median and IQR bias, especially between 5 and 1 km resolutions.

Finally, we also investigated (not shown) the dependence of the median and IQR bias on the MCMC parameters M, C and n discussed in section 3 above. The dependence on these parameters (about the values chosen in section 3) was minimal, illustrating a certain robustness to the MCMC approach. M was varied over 1–5 (control 3), C was varied over 0.0001–1 (control 0.01) and n was varied over 5000–20 000 (control 10 000). Note that there is a danger in going to smaller n values because all nine (or, in section 3, five) truth/initial-condition triangles used in these illustrative studies have the same P_L and Δ – they are just translations in S of the same triangle PDF, the translated PDFs spanning clear to overcast cases. Because of this, errors in P_L, for example, will tend to zero as $n \to 0$, simply because the initial error is zero. Thus, a large number of chain elements n was used just to 'lose track of the initial condition', so to speak. Certainly, a more thorough study would be possible using a large

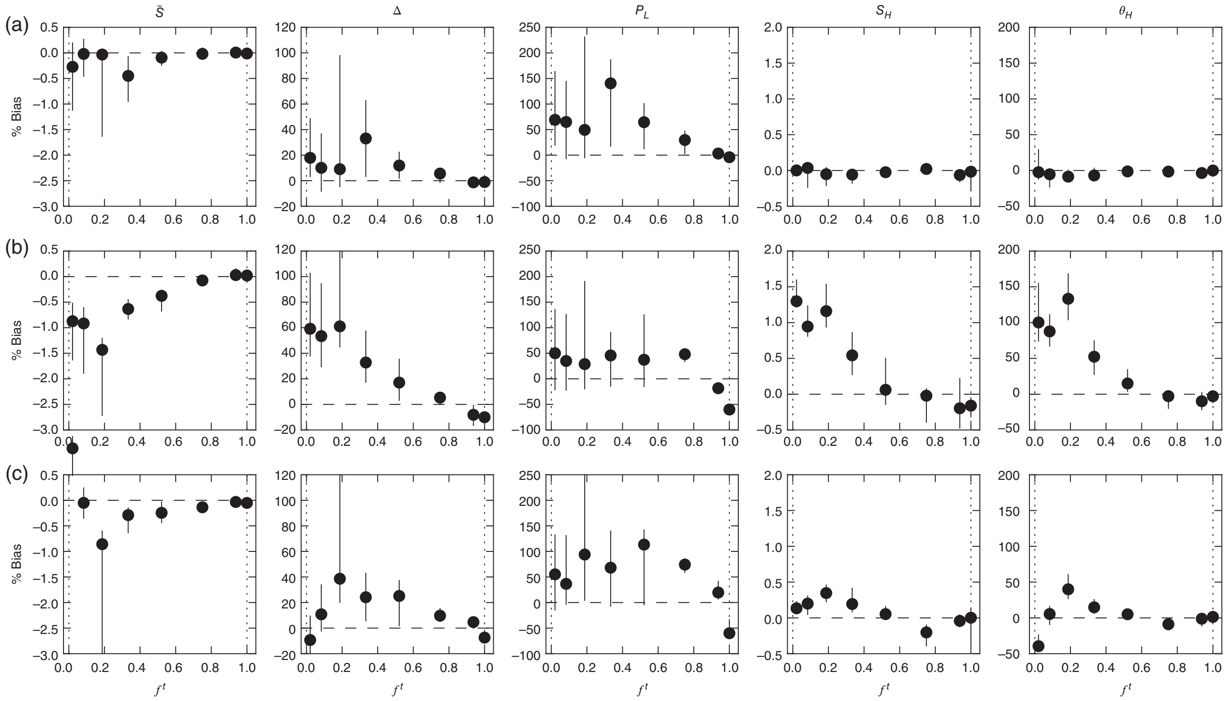

Figure 4. For each row, the percentage biases of the maximum *a posteriori* (MAP) values of five parameters ($\bar{S}$, Δ, $P_{\rm L}$, $S_{\rm H}$ and $\theta_{\rm H}$) from their truth values, as a function of true cloud fraction ($f^{\rm t}$, for eight different truth triangles). The solid circles show the median and the vertical lines the interquartile range, over 90 cases (ten observable realizations for each truth triangle and nine different initial triangles). No prior was used. (a) Top row: the *exact analytic* triangular likelihood model. The first six $f^{\rm t}$ correspond to truth triangles with modes $S_*^t \leq 1$ and have large biases in $P_{\rm L}$ and Δ, since these triangles suffer from the identifiability problem discussed in the text. The biases of the parameters $S_{\rm H}$ and $\theta_{\rm H}$ not affected by this identifiability issue are more acceptable. Note the generally smaller biases for large cloud fractions – these show the value of near-full sampling of the truth triangles by cloudy observations and suggest the potential benefit of adding vapour observations in partially cloudy cases. (b) Middle row: the *approximate KDE-based* likelihood with $N_{\rm sim} = N_{\bullet\max} = 64$. (c) Bottom row: same, but with $N_{\rm sim} = N_{\bullet\max} = 128$. Please refer to the text in section 3.1 for a discussion of the analytic versus KDE comparison.

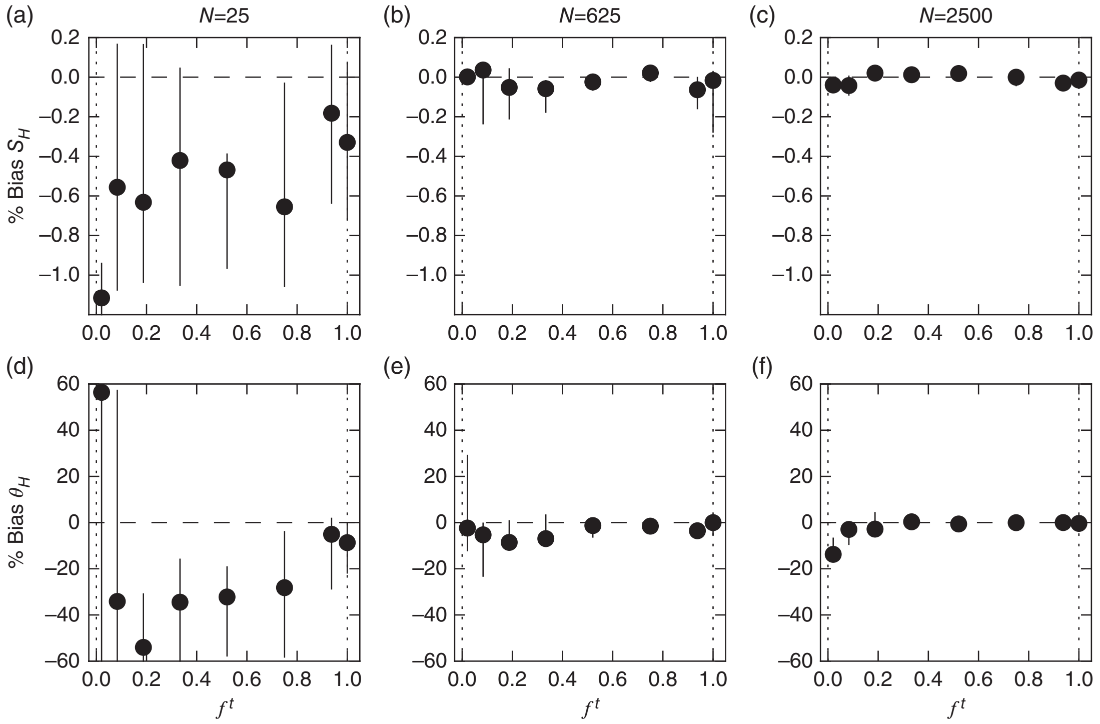

Figure 5. As for Figure 4, but for the analytic likelihood only and for parameters $S_{\rm H}$ and $\theta_{\rm H}$ (rows) and different numbers of observations N (columns). These N are the approximate numbers of pixels in a 1/4° gridbox for 5 km, 1 km and 500 m pixels, respectively. Clearly, increasing the number of observations reduces both the median and interquartile range bias, especially between the 5 and 1 km resolutions.

randomized ensemble of truth/initial-condition triangles with variable $P_{\rm L}$ and Δ as well as $\bar{S}$. However, then the issue becomes selecting a representative distribution of these triangles, since that distribution will vary with latitude, synoptic condition, height in the boundary layer/free atmosphere, etc. Instead, we defer any further testing to Part 2 of this series, where we investigate the behaviour of the full gridcolumn Monte Carlo algorithm, assimilating actual MODIS cloud data and with contemporaneous backgrounds from a GEOS-5 analysis.

4. Summary and discussion

This completes the essential description of the method of Monte Carlo Bayesian cloud data assimilation we are using.

The method was designed with the goal of addressing several common problems in cloud data assimilation: (i) the mismatch between the frequently small scales of cloud variability and typical GCM-gridcolumn footprints is handled by a detailed sub-gridcolumn model of moisture variability, including layer moisture PDFs and vertical coupling of the PDFs using a Gaussian copula model; (ii) the strong nonlinearities present in cloud processes are addressed using a nonlinear, non-gradient parameter space exploration method, Markov chain Monte Carlo (MCMC) Bayesian inference. In particular, a key problem is that a subsaturated background state cannot produce clouds via any small equilibrium perturbation to moisture, but the MCMC approach allows equilibrium jumps into regions of non-zero cloud probability. Another advantage of the MCMC Bayesian approach is that it characterizes the *a posteriori* PDF of control parameters, thereby providing error estimates for the new analyzed state.

The ultimate goal of this work is to produce a fully cycling data assimilation system, one in which the model total water PDF parameters are re-initialized with the values coming from the 'cloud analysis', with the GCM producing a first guess for the next cloud analysis. To achieve this goal, we have structured the project with two distinct milestones: (i) development and assessment of the 'cloud analysis' step by means of MCMC and (ii) update of the PDF scheme in GEOS-5 to provide the time evolution of the triangular PDF parameters and to use the triangular PDF and subcolumn generation consistently in the radiation parametrization and throughout GEOS-5. This article focuses on milestone (i). Moreover, the improved moisture/cloud state afforded by the MCMC algorithm is intended to provide a better background for the hybrid ensemble/4D-Var algorithms in GEOS-5, as well as to assist in the development of proper observation operators for cloudy radiances (taking into consideration cloud overlapping and subgrid variability).

For computational feasibility, we have stayed away from a multivariate cloud analysis involving wind–mass coupling in the MCMC algorithm. Having obtained mass analysis increments, the corresponding wind increments could in principle be generated by using the balance operators in our hybrid grid-point statistical interpolation (GSI) system or relegated altogether to the full meteorological analysis. The tacit goal of our approach is to extract cloud information from the wealth of visible and IR sensors as an intermediate step and to use this information to constrain better the assimilation of IR and microwave cloudy radiances in the main meteorological data assimilation system. What we present here is an incremental step in this direction.

The characterization of observation error and its impact on the performance of our algorithm is also an important aspect that deserves in-depth consideration but has not been addressed in this article. We have recently built a detailed MODIS Cloud Retrieval Simulator (Wind *et al.*, 2013), where a full spectrum of MODIS Level 1 radiances is simulated from subcolumns sampled from triangular PDFs of total water (basically the gridcolumn statistical model of this study) by detailed scattering calculations. These radiances are then fed to the operational MODIS cloud retrieval suite. Besides evaluating cloud retrieval accuracy under a variety of scenarios, this device will allow us to characterize the proper averaging kernels and improve the specification of forward operator error for our cloud assimilation algorithm. Furthermore, these synthetic retrievals can be input to our cloud assimilation algorithm to assess its ability to recover the triangular PDF parameters that were specified in the front end of this simulation chain. This OSSE activity is a project in itself and is beyond the scope of this article.

In section 3, we provided an illustrative testing of the Bayesian MCMC method in a simple one-layer context. These tests illustrated the basic success of the method in reconstructing reasonable analysis triangles using synthetic censored condensate observations, even starting from distant initial conditions with low or zero posterior probability. The testing also revealed some identifiability problems for certain cloudiness regimes, although these problems are expected to disappear in multi-layer cases where multiple triangles are combined with a realistic degree of vertical decorrelation. Discussion of the performance of the new method in the full multi-layered context is presented in Part 2, where the method is validated in various ways and its sensitivity to numerous algorithmic and physical parameters is examined.

Appendix A, which follows, deals with important details of the skewed triangle PDF that we use for layer total moisture marginals and, in particular, some non-trivial details associated with initialization of the PDFs from typical GCM gridbox mean variables. Appendices B and C present technical aspects of the all-sky likelihood evaluation and the forward modelling of cloud-top pressure and brightness temperature. Finally, Appendix D provides a detailed mathematical analysis of the perfect sampling solution to the one-layer test cases of section 3.

Acknowledgements

This work was supported by a NASA grant from the Modeling, Analysis, and Prediction program (solicitation NNH08ZDA001N-MAP, proposal title: 'Assimilation of *A-Train* satellite data for constraining a new PDF-based cloud parametrization in GEOS-5', PI: Arlindo da Silva). The authors thank Steven Platnick and Gala Wind for much useful information on the inner workings of the MODIS cloud algorithms and the characteristics of their retrievals. The authors also thank Dr Chris Snyder and an anonymous reviewer for their reviews, which helped to improve this article substantially . Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center.

Appendices

Appendix A: The skewed triangle distribution

Key results are presented below for the skewed triangular distribution. Only the most important derivation details are included. Further details can be obtained from the authors.

A1. Basics

Consider a skewed triangular PDF in a variable S as follows:

$$p_S(s) = \frac{2}{\Delta}\begin{cases} 0, & s \le S_\mathrm{L}, \\ (s - S_\mathrm{L})/\Delta_\mathrm{L}, & S_\mathrm{L} \le s \le S_*, \\ (S_\mathrm{H} - s)/\Delta_\mathrm{H}, & S_* \le s \le S_\mathrm{H}, \\ 0, & s \ge S_\mathrm{H}, \end{cases} \tag{A1}$$

where $\Delta_\mathrm{L} \equiv S_* - S_\mathrm{L} > 0$, $\Delta_\mathrm{H} \equiv S_\mathrm{H} - S_* > 0$ and $\Delta \equiv \Delta_\mathrm{L} + \Delta_\mathrm{H} = S_\mathrm{H} - S_\mathrm{L}$. The triangle has a base $[S_\mathrm{L}, S_\mathrm{H}]$ of length Δ, a mode at S_*, a height of $2/\Delta$ and unit area as required. The area of the lower section (below S_*) is $P_\mathrm{L} \equiv \Delta_\mathrm{L}/\Delta \in (0,1)$ and the area of the upper section (above S_*) is $P_\mathrm{H} \equiv \Delta_\mathrm{H}/\Delta = 1 - P_\mathrm{L}$. The CDF is

$$\begin{aligned} P_S(s) &= \int_{-\infty}^{s} p_S(x)\,\mathrm{d}x \\ &= \begin{cases} 0, & s \le S_\mathrm{L}, \\ P_\mathrm{L}\left(\frac{s - S_\mathrm{L}}{\Delta_\mathrm{L}}\right)^2, & S_\mathrm{L} \le s \le S_*, \\ 1 - P_\mathrm{H}\left(\frac{S_\mathrm{H} - s}{\Delta_\mathrm{H}}\right)^2, & S_* \le s \le S_\mathrm{H}, \\ 1, & s \ge S_\mathrm{H}. \end{cases} \end{aligned} \tag{A2}$$

Therefore, a simple way to generate a random sample S from the distribution is as follows: (i) generate a uniform random number U on $[0,1]$; (ii) if $U \le P_\mathrm{L}$, $S = S_\mathrm{L} + \Delta_\mathrm{L}\sqrt{U/P_\mathrm{L}}$; (iii) otherwise, $S = S_\mathrm{H} - \Delta_\mathrm{H}\sqrt{(1 - U)/P_\mathrm{H}}$.

A2. Mean and variance

The mean is

$$\bar{S} = \int_{S_{\mathrm{L}}}^{S_{\mathrm{H}}} s\, p_S(s)\, \mathrm{d}s = S_* + \delta/3, \tag{A3}$$

where $\delta \equiv \Delta_{\mathrm{H}} - \Delta_{\mathrm{L}}$. The variance is

$$\sigma_S^2 = \int_{S_{\mathrm{L}}}^{S_{\mathrm{H}}} (s - \bar{S})^2\, p_S(s)\, \mathrm{d}s = \left(\Delta_G^2 + \delta^2/3\right)/6, \tag{A4}$$

where $\Delta_G \equiv \sqrt{\Delta_{\mathrm{L}}\Delta_{\mathrm{H}}}$.

A3. Clear/cloudy decomposition

Consider a triangular distribution in the total saturation ratio $S \equiv q_{\mathrm{t}}/q_{\mathrm{s}}$, with the usual definitions (Norris *et al.*, 2008). Then, under the standard bulk assumption (i.e. all water in excess of saturation is assumed to be condensate), the clear $(0 \le S \le 1)$ fraction is just $f' \equiv P_S(1)$ and the cloudy $(S > 1)$ fraction is $f \equiv 1 - f'$.

Let $[S]_\circ$ be the integral of S over the clear part of the gridbox:

$$[S]_\circ \equiv \int_0^1 s\, p_S(s)\, \mathrm{d}s = I_S(1), \tag{A5}$$

where

$$\begin{aligned} I_S(s_0) &\equiv \int_{-\infty}^{s_0} s\, p_S(s)\, \mathrm{d}s = S_* P_S(s_0) \\ &+ \frac{1}{\Delta} \begin{cases} 0, & s_0 \le S_{\mathrm{L}}, \\ s_0'^2 (1 + \frac{2s_0'}{3\Delta_{\mathrm{L}}}) - \Delta_{\mathrm{L}}^2/3, & s_0 \in [S_{\mathrm{L}}, S_*], \\ s_0'^2 (1 - \frac{2s_0'}{3\Delta_{\mathrm{H}}}) - \Delta_{\mathrm{L}}^2/3, & s_0 \in [S_*, S_{\mathrm{H}}], \\ \Delta_{\mathrm{H}}^2/3 - \Delta_{\mathrm{L}}^2/3, & s_0 \ge S_{\mathrm{H}}, \end{cases} \end{aligned} \tag{A6}$$

with $s_0' \equiv s_0 - S_*$.

We ignore temperature variability, so q_{t} is just S scaled by a constant q_{s}. The clear portion of the vapour integral is then just $[q_{\mathrm{v}}]_\circ = [S]_\circ q_{\mathrm{s}}$, since $q_{\mathrm{t}} = q_{\mathrm{v}}$ for $S \le 1$, and the mean vapour content in the clear portion is $\hat{q}_{\mathrm{v}\circ} = [q_{\mathrm{v}}]_\circ/f'$. For the cloudy portion, $q_{\mathrm{v}} = q_{\mathrm{s}}$ and so the cloudy portion mean vapour is just $\hat{q}_{\mathrm{v}\bullet} = q_{\mathrm{s}}$ and the cloudy portion vapour integral $[q_{\mathrm{v}}]_\bullet = f q_{\mathrm{s}}$. Finally, we have $\bar{q}_{\mathrm{t}} = \bar{S} q_{\mathrm{s}} = [q_{\mathrm{v}}]_\circ + [q_{\mathrm{v}}]_\bullet + [q_{\mathrm{c}}]$, where $[q_{\mathrm{c}}] = f\hat{q}_{\mathrm{c}}$ is the gridbox condensate integral and $\hat{q}_{\mathrm{c}}$ is the in-cloud mean condensate content.

A4. Initialization for partially cloudy gridboxes

Say we have a model with a prognostic set of variables $\{f, \hat{q}_{\mathrm{v}\circ}, \hat{q}_{\mathrm{c}}\}$ for each gridbox. These variables may be simply derived from the PDF of S using results from the previous section. Specifically,

$$f' = P_S(1), \qquad f = 1 - f', \qquad \hat{q}_{\mathrm{v}\circ} = [S]_\circ q_{\mathrm{s}}/f' \tag{A7}$$

and

$$\begin{aligned} \hat{q}_{\mathrm{c}} &= [q_{\mathrm{c}}]/f = (\bar{S} q_{\mathrm{s}} - [q_{\mathrm{v}}]_\circ - [q_{\mathrm{v}}]_\bullet)/f \\ &= q_{\mathrm{s}}(\bar{S} - [S]_\circ - f)/f. \end{aligned} \tag{A8}$$

But how about moving in the opposite direction: from $\{f, \hat{q}_{\mathrm{v}\circ}, \hat{q}_{\mathrm{c}}\}$ to a unique PDF?

The clear $(f = 0)$ and overcast $(f = 1)$ cases are special and are discussed in section 2.3.1 of the main text. For the clear case, $\hat{q}_{\mathrm{c}}$ is undefined and there are many possible triangles with the same $\hat{q}_{\mathrm{v}\circ} = \bar{q}_{\mathrm{v}}$. Similarly, for the overcast case, $\hat{q}_{\mathrm{v}\circ}$ is undefined and there are many possible triangles with the same $\hat{q}_{\mathrm{c}} = \bar{q}_{\mathrm{c}}$. These cases are therefore undetermined and require special treatment.

In this section, therefore, we consider only the partially cloudy gridbox, which has $0 < f < 1$ and $S_{\mathrm{L}} < 1 < S_{\mathrm{H}}$. In this case we have three equations – the second and third parts of (A7) plus (A8) –constrained by $\{f, \hat{q}_{\mathrm{v}\circ}, \hat{q}_{\mathrm{c}}\}$ and three unknowns $\{S_{\mathrm{L}}, S_*, S_{\mathrm{H}}\}$ or $\{\bar{S}, \Delta, P_{\mathrm{L}}\}$, so we hope to find a unique solution PDF. Consider two subcases.

(a) $S_* \ge 1$: then, after some algebra, we must solve

$$\begin{aligned} f' = 1 - f &= P_{\mathrm{L}}(1 - \sigma_{\mathrm{L}})^2, \\ 1 - \hat{q}_{\mathrm{v}\circ}/q_{\mathrm{s}} &= \Delta\{-P_{\mathrm{L}}\sigma_{\mathrm{L}} + P_{\mathrm{L}}^2/(3f') \\ &\qquad - P_{\mathrm{L}}^2\sigma_{\mathrm{L}}^2(1 - 2\sigma_{\mathrm{L}}/3)/f'\}, \\ \hat{q}_{\mathrm{c}}/q_{\mathrm{s}} &= \Delta\{+P_{\mathrm{L}}\sigma_{\mathrm{L}} + P_{\mathrm{H}}^2/(3f) \\ &\qquad - P_{\mathrm{L}}^2\sigma_{\mathrm{L}}^2(1 - 2\sigma_{\mathrm{L}}/3)/f\}, \end{aligned} \tag{A9}$$

where $\sigma_{\mathrm{L}} \equiv (S_* - 1)/\Delta_{\mathrm{L}} \in [0, 1)$. Note that

$$\begin{aligned} Q &\equiv f'(\hat{q}_{\mathrm{v}\circ}/q_{\mathrm{s}} - 1) + f(\hat{q}_{\mathrm{c}}/q_{\mathrm{s}}) \\ &= \Delta\left\{P_{\mathrm{L}}\sigma_{\mathrm{L}} + (1 - 2P_{\mathrm{L}})/3\right\}, \end{aligned} \tag{A10}$$

so, combining with the first two parts of (A9),

$$\begin{aligned} R_{\mathrm{L}} &\equiv (1 - \hat{q}_{\mathrm{v}\circ}/q_{\mathrm{s}})/(Qf') \\ &= \frac{1 - \sigma_{\mathrm{L}}}{f'(3\sigma_{\mathrm{L}} - 2) + (1 - \sigma_{\mathrm{L}})^2}. \end{aligned} \tag{A11}$$

This gives a quadratic equation in σ_{L}.

(b) $S_* \le 1$: then, after some algebra, we must solve

$$\begin{aligned} f = 1 - f' &= P_{\mathrm{H}}(1 - \sigma_{\mathrm{H}})^2, \\ 1 - \hat{q}_{\mathrm{v}\circ}/q_{\mathrm{s}} &= \Delta\{+P_{\mathrm{H}}\sigma_{\mathrm{H}} + P_{\mathrm{L}}^2/(3f') \\ &\qquad - P_{\mathrm{H}}^2\sigma_{\mathrm{H}}^2(1 - 2\sigma_{\mathrm{H}}/3)/f'\}, \\ \hat{q}_{\mathrm{c}}/q_{\mathrm{s}} &= \Delta\{-P_{\mathrm{H}}\sigma_{\mathrm{H}} + P_{\mathrm{H}}^2/(3f) \\ &\qquad - P_{\mathrm{H}}^2\sigma_{\mathrm{H}}^2(1 - 2\sigma_{\mathrm{H}}/3)/f\}, \end{aligned} \tag{A12}$$

where $\sigma_{\mathrm{H}} \equiv (1 - S_*)/\Delta_{\mathrm{H}} \in [0, 1)$. Note that

$$\begin{aligned} Q &\equiv f'(\hat{q}_{\mathrm{v}\circ}/q_{\mathrm{s}} - 1) + f(\hat{q}_{\mathrm{c}}/q_{\mathrm{s}}) \\ &= -\Delta\left\{P_{\mathrm{H}}\sigma_{\mathrm{H}} + (1 - 2P_{\mathrm{H}})/3\right\}, \end{aligned} \tag{A13}$$

so, combining with the first and third parts of (A12),

$$\begin{aligned} R_{\mathrm{H}} &\equiv -(\hat{q}_{\mathrm{c}}/q_{\mathrm{s}})/(Qf) \\ &= \frac{1 - \sigma_{\mathrm{H}}}{f(3\sigma_{\mathrm{H}} - 2) + (1 - \sigma_{\mathrm{H}})^2}. \end{aligned} \tag{A14}$$

This gives a quadratic equation in σ_{H}.

If we introduce the placeholders $\{\mathcal{P}, R, \mathfrak{f}, \sigma\}$ to represent $\{P_{\mathrm{H}}, R_{\mathrm{H}}, f, \sigma_{\mathrm{H}}\}$ for $S_* \le 1$ and $\{P_{\mathrm{L}}, R_{\mathrm{L}}, f', \sigma_{\mathrm{L}}\}$ for $S_* \ge 1$, then the first parts of equations (A9) and (A12) can be combined as

$$\mathfrak{f} = \mathcal{P}(1 - \sigma)^2, \tag{A15}$$

while (A11) and (A14) also have the same form and can be rewritten as

$$\sigma^2 + (3\mathfrak{f} - 2 + R^{-1})\sigma + (1 - 2\mathfrak{f} - R^{-1}) = 0. \tag{A16}$$

This equation can be solved for σ_{L} and σ_{H} (with $\mathfrak{f}$ and R chosen as above) and then P_{L} and P_{H} backed out from (A15). Then Δ can be found via (A10) and (A13) and S_* via the definitions of σ_{L} and σ_{H}, respectively. Also, note that the definitions of Q in (A10) and (A13) are the same and, in fact, Q has a very simple form in terms of the mean total saturation ratio $\bar{S} = \bar{q}_{\mathrm{t}}/q_{\mathrm{s}}$, namely

$$\begin{aligned} Q &= f'(\hat{q}_{\mathrm{v}\circ}/q_{\mathrm{s}} - 1) + f(\hat{q}_{\mathrm{c}}/q_{\mathrm{s}}) \\ &= \bar{q}_{\mathrm{v}}/q_{\mathrm{s}} - f - f' + \bar{q}_{\mathrm{c}}/q_{\mathrm{s}} = \bar{S} - 1, \end{aligned} \tag{A17}$$

since $\bar{q}_{\mathrm{c}} = \hat{q}_{\mathrm{c}} f$ and $\bar{q}_{\mathrm{v}} = \hat{q}_{\mathrm{v}\circ} f' + q_{\mathrm{s}} f$.

After some analysis, the only solution to (A16) for $\mathfrak{f} \in (0,1)$ and $\sigma \in [0,1)$ and $\mathcal{P} \in (0,1)$ is

$$\sigma = 1 - \tau - \sqrt{\tau^2 - \mathfrak{f}}, \tag{A18}$$

where $\tau = \mathfrak{f}/2 \cdot [(3-\mathfrak{f}) + (1-\mathfrak{f})\mathfrak{R}]$, $\mathfrak{R} = \mathcal{R}$ for $S_* \geq 1$ and $\mathfrak{R} = \mathcal{R}^{-1}$ for $S_* \leq 1$, and

$$\mathcal{R} \equiv (\hat{q}_{\mathrm{c}}/q_{\mathrm{s}})/(1 - \hat{q}_{\mathrm{v}\circ}/q_{\mathrm{s}}). \tag{A19}$$

Furthermore, (A18) is only valid for $\mathfrak{f} \in (F^{-1}(\mathfrak{R}), (1+\mathfrak{R})^{-1}]$, where

$$F^{-1}(x) \equiv [\sqrt{1 + 8/(x+1)} - 1]^2/4, \quad \forall\, x > 0. \tag{A20}$$

This means that $f \in (F^{-1}(1/\mathcal{R}), 1/(1+1/\mathcal{R})]$ for $S_* \leq 1$ and $f' \in (F^{-1}(\mathcal{R}), 1/(1+\mathcal{R})]$ for $S_* \geq 1$, or, alternatively, $f \in (f_{\mathrm{lo}}, f_{\mathrm{md}}]$ for $S_* \leq 1$ and $f \in [f_{\mathrm{md}}, f_{\mathrm{hi}})$ for $S_* \geq 1$, where

$$\begin{aligned} f_{\mathrm{lo}} &\equiv F^{-1}(1/\mathcal{R}), \\ f_{\mathrm{md}} &\equiv \mathcal{R}/(1+\mathcal{R}) \text{ and} \\ f_{\mathrm{hi}} &\equiv 1 - F^{-1}(\mathcal{R}). \end{aligned} \tag{A21}$$

For $\mathcal{R} = 1$, these f ranges are $(0.382, 0.5]_{S_* \leq 1}$ and $[0.5, 0.618)_{S_* \geq 1}$. Figure A1 shows a graphical illustration of the solution space. Evidently, there is only a relatively small range of f for which a skewed triangle PDF solution exists for a given $\mathcal{R}$. Note that (i) panels (b) and (c) use a scaled cloud fraction $f^S \equiv (f - f_{\mathrm{lo}})/(f_{\mathrm{hi}} - f_{\mathrm{lo}})$ to expand out the behaviour of σ and P_{L} on $(f_{\mathrm{lo}}, f_{\mathrm{hi}})$ and (ii) $f = f_{\mathrm{md}}(\mathcal{R})$ corresponds to $S_* = 1$ and therefore $\sigma = 0$ and $\mathcal{P} = \mathfrak{f}$, i.e. $P_{\mathrm{L}} = 1 - f$.

A4.1. A reduced P_{L} range

Rather than considering the full $P_{\mathrm{L}} \in (0,1)$ range, we consider only a reduced range $P_{\mathrm{L}} \in I_{\delta_P} \equiv [\delta_P, 1-\delta_P]$, where $\delta_P \in (0, 0.5)$ and typically $\delta_P^2 \ll 1$. We use $\delta_P = 0.1$ for the results in this article. The reason is that $P_{\mathrm{L,H}} \to 0$ represents extreme skewness cases that are usually not found in nature. For example, the region between the $P_{\mathrm{L}} = 0.05$ (dark blue) and $P_{\mathrm{L}} = 0.95$ (brown) contours of Figure A1(c) indicates the subset of $(\mathcal{R}, f^S)$ space for the $\delta_P = 0.05$ case. It is evident from the figure that the range $P_{\mathrm{L}} \in I_{\delta_P}$ will be confined to $S_* \leq 1$ for small enough $\mathcal{R}$ and $S_* \geq 1$ for large enough $\mathcal{R}$, with intermediate $\mathcal{R}$ including both branches. To explore this, imagine Figure A1(c) replotted against f not f^S, since the latter is simply a scaling, dependent only on $\mathcal{R}$, to make visualization easier. A contour $P_{\mathrm{L}}(\mathcal{R}, f) = p$ on this figure will be denoted by the line $f_{P_{\mathrm{L}}=p}(\mathcal{R})$. Each such contour crosses the magenta dashed line, representing the transition $f_{\mathrm{md}}(\mathcal{R}) = (1 + 1/\mathcal{R})^{-1}$ between $S_* \leq 1$ below and $S_* \geq 1$ above. As noted earlier, on this line $S_* = 1$ and therefore $P_{\mathrm{L}} = 1 - f$. Consequently, the intersection of the contour p and the transition line has $f = 1 - p$. Let the $\mathcal{R}$ coordinate at this intersection be denoted $\mathcal{R}^*_{P_{\mathrm{L}}=p}$. Then $(1 + 1/\mathcal{R}^*_{P_{\mathrm{L}}=p})^{-1} = 1 - p$ or

$$\mathcal{R}^*_{P_{\mathrm{L}}=p} = (1-p)/p. \tag{A22}$$

Note the symmetry $\mathcal{R}^*_{P_{\mathrm{L}}=1-p} = p/(1-p) = 1/\mathcal{R}^*_{P_{\mathrm{L}}=p}$. Now, let $\mathcal{R}_{\mathrm{hi}}(\delta_P) \equiv \mathcal{R}^*_{P_{\mathrm{L}}=\delta_P} = (1-\delta_P)/\delta_P$ and $\mathcal{R}_{\mathrm{lo}}(\delta_P) \equiv \mathcal{R}^*_{P_{\mathrm{L}}=1-\delta_P} = 1/\mathcal{R}_{\mathrm{hi}}(\delta_P)$. (For example, $\mathcal{R}_{\mathrm{hi}}(0.2) = 0.8/0.2 = 4$ and $\mathcal{R}_{\mathrm{lo}}(0.2) = 0.2/0.8 = 0.25$, which can be verified in Figure A1(c) as the $\mathcal{R}$ at the intersection of the transition line with the $P_{\mathrm{L}} = 0.2$ (blue) and 0.8 (red) contours, respectively.) Now, with this notation, it is clear from the figure that if $\mathcal{R} \leq \mathcal{R}_{\mathrm{lo}}(\delta_P)$ then $P_{\mathrm{L}} \in I_{\delta_P}$ is confined to $S_* \leq 1$ and if $\mathcal{R} \geq \mathcal{R}_{\mathrm{hi}}(\delta_P)$ then $P_{\mathrm{L}} \in I_{\delta_P}$ is confined to $S_* \geq 1$, while for $\mathcal{R} \in (\mathcal{R}_{\mathrm{lo}}(\delta_P), \mathcal{R}_{\mathrm{hi}}(\delta_P))$, $P_{\mathrm{L}} \in I_{\delta_P}$ spans both branches.

To proceed, consider the solution $\mathfrak{f}_{\mathcal{P}=p}(\mathfrak{R})$ that solves $\mathcal{P}(\mathfrak{R}, \mathfrak{f}) = p \in (0,1)$. This is a contour p of $\mathcal{P}(\mathfrak{R}, \mathfrak{f})$ and is related to the contours of $P_{\mathrm{L}}(\mathcal{R}, f^S)$ in Figure A1(c). From this figure and the definitions of $\mathfrak{f}$ and $\mathfrak{R}$ in the two S_* branches, it is clear that each $\mathcal{P}$ contour can be followed in decreasing $\mathfrak{R}$ till it intersects with the $\mathfrak{f}_{\mathrm{max}}(\mathfrak{R}) = (1+\mathfrak{R})^{-1}$ line, for which $\sigma = 0$ and therefore $\mathcal{P} = \mathfrak{f} = (1+\mathfrak{R})^{-1}$. This gives a minimum possible $\mathfrak{R}$ at this intersection of $\mathfrak{R}_{\mathrm{min}}(p) = (1-p)/p$. Using (A15) and (A18) and its derivation details, it can be shown that $\mathcal{P}(\mathfrak{R}, \mathfrak{f}) = p$ yields a monic trinomial in $\sqrt{\mathfrak{f}}$, with coefficients depending on p and $\mathfrak{R}$, finally yielding

$$\begin{aligned} \mathfrak{f}_{\mathcal{P}=p}(\mathfrak{R}) &= 4\psi(\mathfrak{R}) \\ &\quad \times \{\cos[\arccos(-X)/3] - \cos[\arccos(+X)/3]\}^2, \end{aligned} \tag{A23}$$

where

$$\begin{aligned} \psi(\mathfrak{R}) &\equiv (1 + \mathfrak{R}/3)/(1+\mathfrak{R}), \\ X &\equiv -\frac{\kappa_p/\sqrt{\psi(\mathfrak{R})}}{2(1+\mathfrak{R}/3)} \text{ and} \\ \kappa_p &\equiv \sqrt{p} + 1/\sqrt{p}. \end{aligned} \tag{A24}$$

Finally, let us apply this to the reduced bounds on f, denoted $f_{\mathrm{lo}}(\delta_P)$ and $f_{\mathrm{hi}}(\delta_P)$, that result from confining P_{L} to I_{δ_P}. Clearly these bounds fall within the outer bounds that apply for $\delta_P = 0$, namely the f_{lo} and f_{hi} of (A21). Now, from our earlier analysis,

1. If $\mathcal{R} \leq \mathcal{R}_{\mathrm{lo}}(\delta_P)$ then $P_{\mathrm{L}} \in I_{\delta_P}$ is confined to $S_* \leq 1$ and so $f_{\mathrm{lo}}(\delta_P) = \mathfrak{f}_{\mathcal{P}=1-\delta_P}(1/\mathcal{R})$ and $f_{\mathrm{hi}}(\delta_P) = \mathfrak{f}_{\mathcal{P}=\delta_P}(1/\mathcal{R})$.
2. If $\mathcal{R} \geq \mathcal{R}_{\mathrm{hi}}(\delta_P)$ then $P_{\mathrm{L}} \in I_{\delta_P}$ is confined to $S_* \geq 1$ and so $f_{\mathrm{lo}}(\delta_P) = 1 - \mathfrak{f}_{\mathcal{P}=\delta_P}(\mathcal{R})$ and $f_{\mathrm{hi}}(\delta_P) = 1 - \mathfrak{f}_{\mathcal{P}=1-\delta_P}(\mathcal{R})$.
3. If $\mathcal{R} \in (\mathcal{R}_{\mathrm{lo}}(\delta_P), \mathcal{R}_{\mathrm{hi}}(\delta_P))$, then $P_{\mathrm{L}} \in I_{\delta_P}$ spans both branches and so $f_{\mathrm{lo}}(\delta_P) = \mathfrak{f}_{\mathcal{P}=1-\delta_P}(1/\mathcal{R})$ and $f_{\mathrm{hi}}(\delta_P) = 1 - \mathfrak{f}_{\mathcal{P}=1-\delta_P}(\mathcal{R})$.

A4.2. Adjusting the range of f for constant $\bar{q}_{\mathrm{v}}$ and $\bar{q}_{\mathrm{c}}$

Thus far in this section, we have obtained the criteria for which a valid skewed triangle PDF may be diagnosed based on specification of $\{f, \hat{q}_{\mathrm{v}\circ}, \hat{q}_{\mathrm{c}}\}$. Specifically, equation (A18), together with (A15) for P_{L} and P_{H}, (A10) and (A13) for Δ and the definitions of σ_{L} and σ_{H} for S_* lead to a full specification of the diagnosed triangle for the partially cloudy case. For a given $\hat{q}_{\mathrm{v}\circ}$ and $\hat{q}_{\mathrm{c}}$, however, we have found that only a narrow range of cloud fractions f, as detailed in (A21) and Figure A1, permit a valid partially cloudy solution, namely one with $\sigma \in [0,1)$ and $P_{\mathrm{L}} \in (0,1)$. If, in addition, we restrict P_{L} to a more physical range, say $[\delta_P, 1-\delta_P]$, then an even more restricted range of f is required, as detailed in section A4.1.

How do we proceed if the cloud fraction f is outside this narrow range? This can presumably happen if we are presented with an f, by either observations or model-based simulations, for which the underlying q_{t} PDF is not a skewed triangle. One reasonable strategy is to solve for cases with in-range f as above, but, for cases with out-of-range f, to first clamp the f value to the respective end value of the valid f range. However, this introduces another problem: if we adjust the value of f while keeping the original $\hat{q}_{\mathrm{v}\circ}$ and $\hat{q}_{\mathrm{c}}$ fixed, then the gridbox means $\bar{q}_{\mathrm{v}}$ and $\bar{q}_{\mathrm{c}}$ will change and these, at least in the case of model input, are more likely to be reliable than $\hat{q}_{\mathrm{v}\circ}$ and $\hat{q}_{\mathrm{c}}$. On the other hand, if we hold $\bar{q}_{\mathrm{v}}$ and $\bar{q}_{\mathrm{c}}$ fixed and change f, then $\hat{q}_{\mathrm{v}\circ}$ and $\hat{q}_{\mathrm{c}}$ will change and therefore $\mathcal{R}$ will change, which in turn will alter the acceptable f bounds.

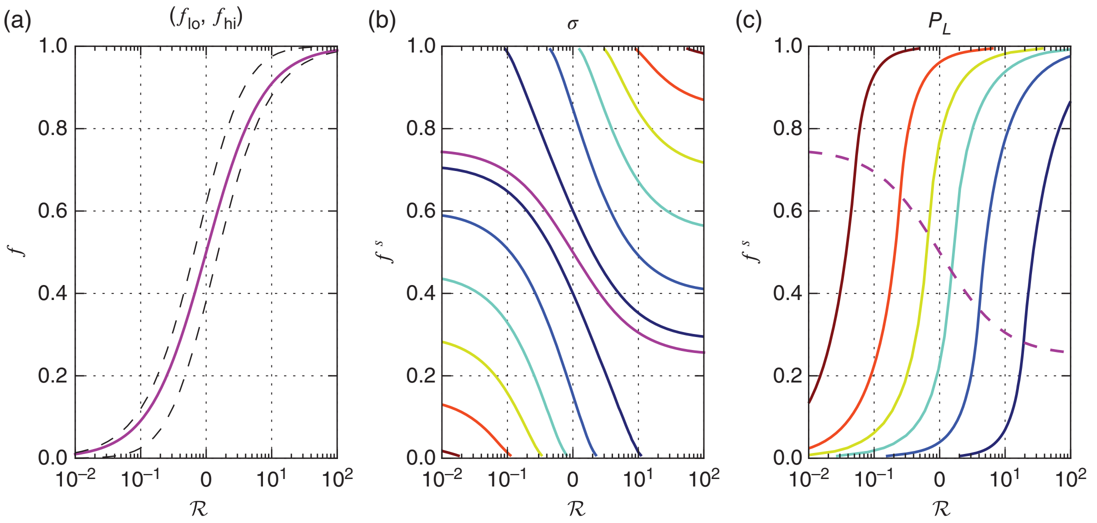

Figure A1. (a) The lower and upper bounds ($f_{\rm lo}, f_{\rm hi}$) on cloud fraction (dashed) for a valid solution, as a function of $\mathcal{R}$, and the transition $f_{\rm md}(\mathcal{R})$ (magenta solid) between $S_* \le 1$ below and $S_* \ge 1$ above, per (A21). (b) The σ solution from (A18) as a function of f and $\mathcal{R}$. The vertical axis uses the normalized cloud fraction $f^S \equiv (f - f_{\rm lo})/(f_{\rm hi} - f_{\rm lo})$. Contours are 0.05 (dark blue), 0.2 (blue), 0.4 (cyan) 0.6 (yellow), 0.8 (red) and 0.95 (brown). The magenta line again shows $S_* = 1 \implies \sigma = 0$. (c) The same, but for $P_{\rm L}$ and with $f^S_{\rm md}(\mathcal{R})$ as the magenta dashed line.

Using (A19) and $\hat{q}_{{\rm v}\circ}/q_{\rm s} = (\bar{q}_{\rm v}/q_{\rm s} - f)/(1-f)$,

$$\begin{aligned} \mathcal{R} &= \mathcal{R}_m \left(f^{-1} - 1\right) \equiv \mathcal{R}(f; \mathcal{R}_m), \\ \mathcal{R}_m &\equiv \frac{\bar{q}_{\rm c}/q_{\rm s}}{1 - \bar{q}_{\rm v}/q_{\rm s}}. \end{aligned} \qquad \text{(A25)}$$

This form emphasizes that $\mathcal{R}$ becomes a function of f if $\bar{q}_{\rm v}$ and $\bar{q}_{\rm c}$ are to be held fixed, rather than $\hat{q}_{\rm c}$ and $\hat{q}_{{\rm v}\circ}$. Note that $\mathcal{R}(f; \mathcal{R}_m)$ can be regarded as a contour of $\mathcal{R}_m$ in $(\mathcal{R}, f)$ space.

Consider a case characterized by $\mathcal{R}_m$ and a cloud fraction f_0. We evaluate $f^0_{\rm lo} \equiv f_{\rm lo}(\mathcal{R}(f_0; \mathcal{R}_m); \delta_P)$ and $f^0_{\rm hi} \equiv f_{\rm hi}(\mathcal{R}(f_0; \mathcal{R}_m); \delta_P)$ using section A4.1 and say that f_0 is outside $[f^0_{\rm lo}, f^0_{\rm hi}]$. First, consider the case where $f_0 < f^0_{\rm lo}$. In $(\mathcal{R}, f)$ space, as in Figure A1, this corresponds to a point $(\mathcal{R}(f_0; \mathcal{R}_m), f_0)$ falling below the contour $f_{\rm lo}(\mathcal{R}; \delta_P)$ on which $P_{\rm L} = \delta_P$. The goal is to increase f from f_0, following the contour $\mathcal{R}(f; \mathcal{R}_m)$ passing though the point, until it intersects the contour $f_{\rm lo}(\mathcal{R}; \delta_P)$. The contour $\mathcal{R}(f; \mathcal{R}_m)$ has decreasing $\mathcal{R}$ for increasing f and so moves up and to the left in Figure A1(a). However, in the process of this move, the contour $f_{\rm lo}(\mathcal{R}; \delta_P)$ moves down and to the left, towards decreasing $f_{\rm lo}$. The intersection point we seek therefore has $f \in (f_0, f^0_{\rm lo})$. Conversely, for $f_0 > f^0_{\rm hi}$, we follow the contour $\mathcal{R}(f; \mathcal{R}_m)$ passing through $(\mathcal{R}(f_0; \mathcal{R}_m), f_0)$, decreasing in f and increasing in $\mathcal{R}$, until it intersects with the contour $f_{\rm hi}(\mathcal{R}; \delta_P)$ on which $P_{\rm L} = 1 - \delta_P$, which has increasing $f_{\rm hi}$ for increasing $\mathcal{R}$. Thus, in this case, the intersection point we seek has $f \in (f^0_{\rm hi}, f_0)$.

In this way, we are able to bound an acceptable solution f for an initially out-of-bounds f_0. In practice, we solve the problem numerically using a simple bisection algorithm. We start by bisecting the appropriate f range above. Of course, we do not know the solution f, but if the new point f_1 is within a small tolerance of the respective $[f^1_{\rm lo}, f^1_{\rm hi}]$ end-point then we consider the solution found. (Here, the 'end-point' is actually an f that is 1% inside that range, near $f^1_{\rm lo}$ if f_0 was initially too low and near $f^1_{\rm hi}$ if f_0 was initially too high; 'small tolerance' means 1% of $f^1_{\rm hi} - f^1_{\rm lo}$.) If not within tolerance, we narrow the f bounds and repeat for a new bisection point f_2 of the updated bounds. (The 'narrowed bounds' are f_1 and the 'end-point' above and are clipped so that the new range never falls outside the previous range.) We continue this iteration until either a solution is found or 100 iterations are performed. If a valid solution is not found or if the solution has $S_{\rm L} < 0$ or $S_{\rm H} > S_{\rm max}$, as discussed in the next section, then we use the fallback solution presented in that section. As a final note, this procedure seems to work well for f_0 clipped to $[0.001, 0.999]$ and for $\mathcal{R}(f_n)$ clipped to $[10^{-6}, 10^6]$.

A4.3. Physical bounds on $S_{\rm L}$ and $S_{\rm H}$

None of our analysis in section A4 has ensured the physical constraint $S_{\rm L} \ge 0$, preventing negative moisture, or the physically reasonable bound constraint $S_{\rm H} \le S_{\rm max}$, preventing excessive moisture. An analytic treatment of these additional constraints has thus far been too complicated to complete, so if these bounds are violated we use the fallback solution of a triangle with base $[0, S_{\rm max}]$, but adjusted to honour $\bar{S}$, as per section A6.

A5. Attempted symmetric triangle solution

For clear or overcast gridbox, we employ a symmetric triangular solution ($P_{\rm L} = 0.5$) with a prescribed width Δ_0 (set to 0.4 for the results in this article, unless otherwise specified). Together with the mean $\bar{S}$, which is the same as the mode S_* for a symmetric triangle, this specifies a unique triangular PDF. However, there is an additional restriction, namely that the triangle must fall wholly within a domain $[S_0, S_1]$, which for clear gridboxes is $[0, 1]$ and for overcast gridboxes is $[1, S_{\rm max}]$, where $S_{\rm max}$ is some reasonable upper bound on allowable total saturation ratio S. (Note that the mean $\bar{S}$ is required to fall in (S_0, S_1) for the clear and overcast cases. The model $\bar{S}$ is clipped to $S_{\rm max}$ if it exceeds this.)

The base of this nominal symmetric triangle, $[\bar{S} - \Delta_0/2, \bar{S} + \Delta_0/2]$, should fit between the bounds S_0 and S_1. If not, one of the following cases applies. For each of these cases, we insist additionally on a $P_{\rm L}$ in the reduced range $[\delta_P, 1 - \delta_P]$, where $\delta_P \in (0, 0.5)$, nominally 0.1 for the results in this article. The cases are as follows.

(a) Both bounds S_0 and S_1 are violated, i.e. $S_{\rm L} = \bar{S} - \Delta_0/2 < S_0$ and $S_{\rm H} = \bar{S} + \Delta_0/2 > S_1$. In this case. we abandon the symmetry restriction and attempt to force the base to exactly $[S_0, S_1]$, if we can do so while still honouring the mean $\bar{S}$. We discuss this case in section A6.

(b) Only S_0 is violated, i.e. $S_{\rm L} = \bar{S} - \Delta_0/2 < S_0$ and $S_{\rm H} = \bar{S} + \Delta_0/2 \le S_1$. In this case, we increase $S_{\rm L}$ to S_0 while holding $\Delta_{\rm H} \equiv S_{\rm H} - S_*$ fixed at $\Delta_0/2$ and $\bar{S}$ fixed at its specified value. Note that, for a *general* triangular PDF, from (A3),

$$S_* = [3\bar{S} - S_{\rm L} - \Delta_{\rm H}]/2. \qquad \text{(A26)}$$

Clearly, as $S_{\rm L}$ is increased, S_* and therefore $S_{\rm H}$ must decrease. This new triangle therefore falls wholly in $[S_0, S_1]$.

The new triangle has

$$P_{\rm L}=\frac{3(\bar{S}-S_0)-\Delta_0/2}{3(\bar{S}-S_0)+\Delta_0/2}<1/2. \tag{A27}$$

To ensure that $P_{\rm L}\ge\delta_P$ requires

$$\bar{S}-S_0\ge\frac{\Delta_0}{6}\frac{1+\delta_P}{1-\delta_P}. \tag{A28}$$

If this is the case, then the new triangle is an acceptable solution, with $P_{\rm L}$ as above and

$$\Delta=\frac{\Delta_0/2}{1-P_{\rm L}}. \tag{A29}$$

If (A28) is violated, we select another new triangle with $S_{\rm L}=S_0$, $P_{\rm L}=\delta_P$ and $\bar{S}$ as specified. Then

$$\Delta=\frac{3(\bar{S}-S_0)}{1+\delta_P}, \tag{A30}$$

completes the specification. (Note that $\Delta_{\rm H}=(1-P_{\rm L})\Delta=3(\bar{S}-S_0)(1-\delta_P)/(1+\delta_P)<\Delta_0/2$, since (A28) is violated. Then, from (A26), $S_{\rm H}=S_*+\Delta_{\rm H}=\bar{S}+[\bar{S}-S_0+\Delta_{\rm H}]/2<\bar{S}+[\bar{S}-S_0+\Delta_0/2]/2$. By the definition of case (b), $\bar{S}-S_0<\Delta_0/2$, so $S_{\rm H}<\bar{S}+\Delta_0/2\le S_1$. This ensures that this triangle has a base in $[S_0,S_1]$, as required.)

(c) Only S_1 is violated, i.e. $S_{\rm L}=\bar{S}-\Delta_0/2\ge S_0$ and $S_{\rm H}=\bar{S}+\Delta_0/2>S_1$. Then, by a very similar analysis to that in case (b), first we try $S_{\rm H}=S_1$, $\Delta_{\rm L}=\Delta_0/2$ and $\bar{S}$ as specified. This triangle has

$$P_{\rm H}=1-P_{\rm L}=\frac{3(S_1-\bar{S})-\Delta_0/2}{3(S_1-\bar{S})+\Delta_0/2}<1/2. \tag{A31}$$

To ensure that $P_{\rm H}\ge\delta_P$ requires

$$S_1-\bar{S}\ge\frac{\Delta_0}{6}\frac{1+\delta_P}{1-\delta_P}. \tag{A32}$$

If this is the case, then the new triangle is an acceptable solution, with $P_{\rm H}$ as above and

$$\Delta=\frac{\Delta_0/2}{1-P_{\rm H}}. \tag{A33}$$

If (A32) is violated, we use the triangle with $S_{\rm H}=S_1$, $P_{\rm H}=\delta_P$ and $\bar{S}$ as specified. Then

$$\Delta=\frac{3(S_1-\bar{S})}{1+\delta_P}, \tag{A34}$$

completes the specification.

A6. End-points specified triangle solution

In this case, we are provided with a range (S_0,S_1) containing $\bar{S}$ and want to choose a triangle with that exact base and $P_{\rm L}$ in the reduced range $[\delta_P,1-\delta_P]$, $\delta_P\in(0,1/2)$. The problem is that only such triangles with

$$\bar{S}\in[S_0+\Delta_*/3,S_1-\Delta_*/3], \tag{A35}$$

where $\Delta_*\equiv(S_1-S_0)(1+\delta_P)$, are possible, since the mean simply cannot take on values too close to the end-points. For $\bar{S}$ in this range, the triangle with base (S_0,S_1) and mean $\bar{S}$ is acceptable and has $\Delta=S_1-S_0$ and $P_{\rm L}=(S_*-S_0)/\Delta$, where $S_*=3\bar{S}-(S_0+S_1)$. For $\bar{S}$ below the lower limit, we reduce $S_{\rm H}$ while holding $S_{\rm L}=S_0$, $P_{\rm L}=\delta_P$ and $\bar{S}$ at the specified value. This gives the reduced base

$$\Delta=\frac{3(\bar{S}-S_0)}{1+\delta_P}. \tag{A36}$$

For $\bar{S}$ above the upper limit, we increase $S_{\rm L}$ while holding $S_{\rm H}=S_1$, $P_{\rm H}=\delta_P$ and $\bar{S}$ at the specified value. This gives the reduced base

$$\Delta=\frac{3(S_1-\bar{S})}{1+\delta_P}. \tag{A37}$$

Appendix B: Bayes inference for clear and cloudy pixels

B1. Form of the likelihood

This Appendix is a justification of the form of the pixel likelihood $\hat{p}(\hat{\boldsymbol{y}}|\boldsymbol{\alpha})$ introduced in section 2.6. Say $\boldsymbol{\alpha}$ is a gridcolumn state vector and $\boldsymbol{y}$ is a vector of gridcolumn observations (comprising multiple pixels and multiple properties per pixel). An underlined version of these or any other quantity will denote a random variable in the quantity. Then, in terms of conditional probabilities,

$$\begin{aligned} &P(\underline{\boldsymbol{\alpha}}\in R_{\boldsymbol{\alpha}}\cap\underline{\boldsymbol{y}}\in R_{\boldsymbol{y}})\\ &\quad=P(\underline{\boldsymbol{\alpha}}\in R_{\boldsymbol{\alpha}}|\underline{\boldsymbol{y}}\in R_{\boldsymbol{y}})\,P(\underline{\boldsymbol{y}}\in R_{\boldsymbol{y}})\\ &\quad=P(\underline{\boldsymbol{y}}\in R_{\boldsymbol{y}}|\underline{\boldsymbol{\alpha}}\in R_{\boldsymbol{\alpha}})\,P(\underline{\boldsymbol{\alpha}}\in R_{\boldsymbol{\alpha}})\\ \Longrightarrow\ &p(\boldsymbol{\alpha}|\boldsymbol{y})\,\cancel{\mathrm{d}\boldsymbol{\alpha}}=\frac{P(\underline{\boldsymbol{y}}\in R_{\boldsymbol{y}}|\boldsymbol{\alpha})\,p(\boldsymbol{\alpha})\,\cancel{\mathrm{d}\boldsymbol{\alpha}}}{P(\underline{\boldsymbol{y}}\in R_{\boldsymbol{y}})}, \end{aligned} \tag{B1}$$

where $R_{\boldsymbol{\alpha}}$ is an infinitesimal region of volume $\mathrm{d}\boldsymbol{\alpha}$ containing $\boldsymbol{\alpha}$ and similarly for $\boldsymbol{y}$ and where $p(\boldsymbol{\alpha}|\boldsymbol{y})$ and $p(\boldsymbol{\alpha})$ are probability densities with respect to $\boldsymbol{\alpha}$.

For Bayesian inference on $\boldsymbol{\alpha}$, we may ignore the denominator (which is invariant to $\boldsymbol{\alpha}$), yielding

$$p(\boldsymbol{\alpha}|\boldsymbol{y})\propto P(\underline{\boldsymbol{y}}\in R_{\boldsymbol{y}}|\boldsymbol{\alpha})\,p(\boldsymbol{\alpha}). \tag{B2}$$

This is similar to (1), but $P(\underline{\boldsymbol{y}}\in R_{\boldsymbol{y}}|\boldsymbol{\alpha})$ is a pure probability, not a density. As in section 2.6, the observations are decomposed into i.i.d. pixels $\hat{\boldsymbol{y}}_1,\ldots,\hat{\boldsymbol{y}}_N$ and so

$$P(\underline{\boldsymbol{y}}\in R_{\boldsymbol{y}}|\boldsymbol{\alpha})=\prod_{n=1}^{N}\hat{P}(\underline{\hat{\boldsymbol{y}}}\in R_{\hat{\boldsymbol{y}}_n}|\boldsymbol{\alpha}), \tag{B3}$$

where $\hat{P}(\cdot|\boldsymbol{\alpha})$ is the common per pixel likelihood. For each clear pixel, $\hat{\boldsymbol{y}}_n$ is identically zero and $\hat{P}(\underline{\hat{\boldsymbol{y}}}\in R_{\boldsymbol{0}}|\boldsymbol{\alpha})$ is $P_\circ(\boldsymbol{\alpha})$, the finite likelihood of a clear pixel. This is true no matter how small $R_{\boldsymbol{0}}$ becomes, since clear pixels have $\hat{\boldsymbol{y}}_n$ exactly zero. For cloudy pixels, we can write $\hat{P}(\underline{\hat{\boldsymbol{y}}}\in R_{\hat{\boldsymbol{y}}_n}|\boldsymbol{\alpha})=\hat{p}(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha})\,\mathrm{d}\hat{\boldsymbol{y}}_n$ and the term $\mathrm{d}\hat{\boldsymbol{y}}_n$ cancels with an identical term in the corresponding ignored denominator term $\hat{P}(\underline{\hat{\boldsymbol{y}}}\in R_{\hat{\boldsymbol{y}}_n})$. In practice, for cloudy pixels, we write $\hat{p}(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha})=P_\bullet(\boldsymbol{\alpha})\,\hat{p}_\bullet(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha})$, where $P_\bullet(\boldsymbol{\alpha})=1-P_\circ(\boldsymbol{\alpha})$ is the likelihood of a cloudy pixel at $\boldsymbol{\alpha}$ and $\hat{p}_\bullet(\hat{\boldsymbol{y}}|\boldsymbol{\alpha})$ is a likelihood density for cloudy pixels only, so that its integral over all cloudy $\hat{\boldsymbol{y}}$ is one.

In summary, we may use (B2) and (B3), but for $\hat{P}(\underline{\hat{\boldsymbol{y}}}\in R_{\hat{\boldsymbol{y}}_n}|\boldsymbol{\alpha})$ substituting $P_\circ(\boldsymbol{\alpha})$ for clear pixels and $P_\bullet(\boldsymbol{\alpha})\,\hat{p}_\bullet(\hat{\boldsymbol{y}}_n|\boldsymbol{\alpha})$ for cloudy pixels.

B2. Technical details of the likelihood evaluation

The algorithmic evaluation of the total log-likelihood $\mathcal{L}$ of (6) has several important shortcuts. These details are of a more technical nature and so were not included in the main description in section 2.6. The reader should read that section before proceeding.

(1) First, the clear term $\mathcal{L}_\circ$ is evaluated per (7). If there are no clear pixels ($N_\circ=0$), then $\mathcal{L}_\circ=0$. If there are clear pixels ($N_\circ>0$) but $P_\circ(\boldsymbol{\alpha})=0$, indicating zero probability of such an event, then evaluation of $\mathcal{L}$ is terminated and zero probability is returned for (1).

(2) Next, the cloudy term $\mathcal{L}_\bullet$ is evaluated. If there are no cloudy pixels ($N_\bullet=0$), then $\mathcal{L}_\bullet=0$. If there are cloudy pixels ($N_\bullet>0$), then the first line $N_\bullet\ln P_\bullet(\boldsymbol{\alpha})$ of (9) is evaluated. If $P_\bullet(\boldsymbol{\alpha})=0$, then $N_\bullet>0$ is impossible and so again the evaluation of $\mathcal{L}$ is terminated and zero probability is returned for (1).

(3) Next the second line of (9), involving p_c-cloudy pixels, is evaluated. If there are no such pixels ($N_{\bullet p_c} = 0$), then the second line is zero. If $N_{\bullet p_c} > 0$ but $P(p_c|\bullet, \boldsymbol{\alpha}) = 0$, we again have an impossibility and zero probability is returned. However, if $P(p_c|\bullet, \boldsymbol{\alpha}) > 0$, we evaluate $N_{\bullet p_c} \ln P(p_c|\bullet, \boldsymbol{\alpha})$ and move on to the summation term involving $\hat{p}_{\bullet p_c}$, which is discussed separately below.
(4) Finally, the last line of (9), involving T_b-cloudy pixels, is evaluated in a completely analogous manner.

Now, regarding the summation term

$$\sum_{n \in \bullet p_c} \ln \hat{p}_{\bullet p_c}((\ln \tau, p_c)_n | \boldsymbol{\alpha}),$$

because the evaluation of $\hat{p}_{\bullet p_c}$ can be somewhat expensive and because the number of pixels per gridcolumn can be large (≈ 625 1 km pixels for a 1/4° model resolution), if $N_{\bullet p_c}$ exceeds some limit $N_{\bullet \max}$ then a random subset of $n \in \bullet p_c$, comprising only $N_{\bullet \max}$ elements, is used instead and the resulting summation term is scaled up by $N_{\bullet p_c}/N_{\bullet \max}$. Currently, we set $N_{\bullet \max}$ equal to the number of simulated subcolumns (per gridcolumn), N_{sim}. An N_{sim} value of 64 is typical and gives approximately 1–2% accuracy in the simulated cloud fraction. We will address the sensitivity to these parameters in Part 2.

With these algorithmic details and many other computational efficiency improvements, the Monte Carlo Bayesian CDA algorithm described herein has a throughput of about 4 months per day for hourly *Aqua* MODIS assimilation on a 1/2° model grid using 32 Westmere nodes (2.8 GHz clock speed, 12 cores per node) on the Discover cluster at the NASA Center for Climate Simulation (NCCS) at the NASA Goddard Space Flight Center.

B3. Details of the KDE evaluation

Our default method for evaluating the likelihood $\hat{p}_{\bullet p_c}((\ln \tau, p_c)|\boldsymbol{\alpha})$ from the subcolumn-generated cloudy p_c-available pairs $\{(\tau, p_c)\}$ is a kernel density estimate (KDE) with Gaussian kernels, as introduced in section 2.6. This KDE can represent complex, multimodal distributions and is our default method.

The method constructs the PDF from the normalized sum of a set of 2D Gaussians, one centred at each simulated point in $\{(\ln \tau, p_c)\}$. The covariance of each of these Gaussians is fixed and equal to the sample covariance of the whole of $\{(\ln \tau, p_c)\}$ multiplied by the square of a factor $f_S = (1/N^{\text{sim}}_{\bullet p_c})^{1/(d+4)}$, called the Scott's factor, where d is the number of dimensions, here two.

Note that if the gridcolumn is a night-time one, τ is not available. In this case, the 2D $\hat{p}_{\bullet p_c}((\ln \tau, p_c)|\boldsymbol{\alpha})$ above reduces to a 1D $\hat{p}_{\bullet p_c}(p_c|\boldsymbol{\alpha})$ and a 1D KDE is used instead.

Appendix C: Forward modelling of p_c and T_b

C1. CO$_2$-slicing cloud-top pressure p_c

The CO$_2$-slicing cloud-top pressure p_c is forward modelled using the same simple approximation used by the COSP MODIS simulator (Bodas-Salcedo *et al.*, 2011), namely

$$p_c = \frac{1}{\Delta\tau} \int_0^{\Delta\tau} p \, d\tau, \qquad \text{(C1)}$$

where $\Delta\tau = \min(\Delta\tau_{\text{CO}_2}, \tau)$ and $\Delta\tau_{\text{CO}_2} = 1.0 \cdot \mu_{\text{sat}}$ is an effective nadir-adjusted optical depth to which the MODIS CO$_2$-slicing algorithm is considered to see into a cloud. To evaluate p_c for the gridcolumn's vertical grid, we assume that pressure and COT are linearly related in each layer, which is physically reasonable (at least if we assume that moisture properties are approximately vertically uniform within a layer and, especially, that a cloud fills the entire vertical extent of each layer in which it occurs, both of which are reasonable discretization assumptions for thin layers). Then

$$p_c \approx \frac{1}{\Delta\tau} \left(\sum_{k=1}^{K^*-1} \bar{p}_k \tau_k + \bar{p}^* \tau^* \right), \qquad \text{(C2)}$$

where K^* is the first layer, counting from the top ($k = 1$), for which the cumulative COT, $\sum_{k=1}^{K^*} \tau_k \geq \Delta\tau$ and $\bar{p}_k$ is the mean of the edge pressures of layer k. For the final layer K^*, τ^* is only that portion of τ_{K^*} needed to bring the cumulative COT up to $\Delta\tau$ and $\bar{p}^*$ is the mean of the edge pressures of this reduced layer (the final end-point pressure evaluated, remembering that pressure is linear with COT in the layer).

C2. Brightness temperature T_b

T_b is evaluated with an IR-only version of the all-sky brightness temperature calculation used in the COSP ICARUS algorithm, as in the Appendix of Klein and Jakob (1999), hereafter KJ. The details are not particularly important to this article, but here is a brief outline: the cloud infrared emissivity is calculated as

$$\epsilon_k^{\text{cld}} = 1 - \exp(-\tau_k^{\text{IR}}),$$

where τ_k^{IR} is the COT at 10.5 μm and is calculated analogously to visible τ_k, but using the IR rather than visible COT routine of the GEOS-5 code. ϵ_k^{cld} is combined with a water-vapour continuum emissivity ϵ_k^{wv} to produce a total emissivity $\epsilon_k = 1 - (1 - \epsilon_k^{\text{wv}})(1 - \epsilon_k^{\text{cld}})$ and this is used with KJ's equations (A4), (A5) and (A7) to produce a TOA radiance I. Then (A5) is inverted, with I replacing KJ's $f\{T^k\}$, to extract the brightness temperature T_b.

Appendix D: A mathematical analysis of Bayesian inference for a simple skewed triangle model

Consider a continuous random variable S with a probability density function $p_S(s)$. Say that S is 'observed' in some way to yield an observation $Y(S)$. Say that the observation is left-censored, in the sense that S at and below some threshold will yield a particular minimum value of Y. For concreteness, we will consider the observation operator $Y = \max(S - 1, 0)$, so that, for $S \leq 1$, $Y = 0$ and for $S > 1$, $Y = S - 1 > 0$. This is a condensate-like observation operator.

Say we make N independent observations $y_1, \ldots, y_N$ of Y. Say N_0 of these observations have $Y = 0$, N_1 have $Y \in (0, \delta]$, N_2 have $Y \in (\delta, 2\delta]$, etc., where δ is a very small Y interval, approaching zero. Say we do not record any more information about the observations other than the ordered set $\mathcal{N} = \{N_0, N_1, N_2, \ldots\}$, most elements of which are zero for finite N. However, $\sum_{i=0}^{\infty} N_i = N$ always.

Let $\Pr(C)$ denote the probability of some condition C. Then, because of the independence of the observations,

$$\Pr(\mathcal{N}) \to P_0^{N_0} \prod_{i=1}^{\infty} \left(p_Y(y^{(i)}) \times \delta \right)^{N_i} \quad \text{as} \quad \delta \to 0, \qquad \text{(D1)}$$

where

(1) $P_0 = \Pr(Y = 0) = \Pr(S \leq 1) = P_S(1)$ is the probability of being censored to $Y = 0$, with P_S being the CDF of S;
(2) $p_Y(y)$ is the PDF of Y, here evaluated for $y > 0$, for which $p_Y(y) = p_S(y + 1)$; and
(3) $y^{(i)}$ is any y in the ith delta bin (i.e. the bin with N_i observations).

Say we parametrize $p_S(s; \boldsymbol{\alpha})$ by a real vector $\boldsymbol{\alpha}$ of parameters. Within the Bayesian framework, the parameter state $\boldsymbol{\alpha}$ is itself a realization of a random variable $\boldsymbol{A}$, with a prior PDF $p_{\boldsymbol{A}}(\boldsymbol{\alpha})$. Let $\mathrm{d}V$ denote some infinitesimal volume in $\boldsymbol{A}$ phase space containing $\boldsymbol{\alpha}$. Bayes' Theorem states that

$$\begin{aligned}\Pr(\boldsymbol{A}\in \mathrm{d}V \cap \mathcal{N}) &= \Pr(\boldsymbol{A}\in \mathrm{d}V|\mathcal{N})\Pr(\mathcal{N})\\ &= \Pr(\mathcal{N}\,|\boldsymbol{A}\in \mathrm{d}V)\Pr(\boldsymbol{A}\in \mathrm{d}V),\end{aligned} \tag{D2}$$

yielding

$$p_{\boldsymbol{A}}(\boldsymbol{\alpha}|\,\mathcal{N})\cancel{\mathrm{d}V}\,\Pr(\mathcal{N}) = \Pr(\mathcal{N}\,|\boldsymbol{A}\in \mathrm{d}V)\,p_{\boldsymbol{A}}(\boldsymbol{\alpha})\cancel{\mathrm{d}V}, \tag{D3}$$

and hence

$$\begin{aligned}p_{\boldsymbol{A}}(\boldsymbol{\alpha}|\mathcal{N}) &= \frac{\Pr(\mathcal{N}|\boldsymbol{\alpha})\,p_{\boldsymbol{A}}(\boldsymbol{\alpha})}{\Pr(\mathcal{N})}\\ &= \frac{(P_S(1;\boldsymbol{\alpha}))^{N_0}\prod_{i=1}^{\infty}\left(p_S(y^{(i)}+1;\boldsymbol{\alpha})\right)^{N_i}\,p_{\boldsymbol{A}}(\boldsymbol{\alpha})}{(P_S(1))^{N_0}\prod_{i=1}^{\infty}\left(p_S(y^{(i)}+1)\right)^{N_i}}.\end{aligned} \tag{D4}$$

Therefore, up to a constant independent of $\boldsymbol{\alpha}$, we may write

$$\begin{aligned}\ln p_{\boldsymbol{A}}(\boldsymbol{\alpha}|\mathcal{N}) &= N_0\ln P_S(1;\boldsymbol{\alpha})\\ &+\sum_{i=1}^{\infty}N_i\ln p_S(y^{(i)}+1;\boldsymbol{\alpha})+\ln p_{\boldsymbol{A}}(\boldsymbol{\alpha}).\end{aligned} \tag{D5}$$

Henceforth, we will take $p_S(s)$ as the simple skewed triangle model $p_\triangle(s)$, as described in section 2.2, Figure 1 and (A1). There are various sets of parameter triplets $\boldsymbol{\alpha}$ that can be used to specify the skewed triangular PDF fully. One example would be $\boldsymbol{\alpha}=(S_*,\Delta,P_\mathrm{L})$, specifying the the modal S, the triangle base and the $S\le S_*$ probability mass, respectively.

In any case, the skewed triangular PDF and CDF, from (A1) and (A2), can be written as follows:

$$p_\triangle(s;\boldsymbol{\alpha}) = \begin{cases}2\,(s-S_\mathrm{L})/\theta_\mathrm{L}, & s\in[S_\mathrm{L},\,S_*],\\ 2\,(S_\mathrm{H}-s)/\theta_\mathrm{H}, & s\in[S_*,\,S_\mathrm{H}],\\ 0, & \text{otherwise}\end{cases} \tag{D6}$$

and

$$P_\triangle(s;\boldsymbol{\alpha}) = \begin{cases}0, & s\le S_\mathrm{L},\\ (s-S_\mathrm{L})^2/\theta_\mathrm{L}, & s\in[S_\mathrm{L},S_*],\\ 1-(S_\mathrm{H}-s)^2/\theta_\mathrm{H}, & s\in[S_*,S_\mathrm{H}],\\ 1, & s\ge S_\mathrm{H},\end{cases} \tag{D7}$$

where $\theta_\mathrm{L}=P_\mathrm{L}\Delta^2$ and $\theta_\mathrm{H}=P_\mathrm{H}\Delta^2$. We call $f'(\boldsymbol{\alpha})=P_\triangle(1;\boldsymbol{\alpha})$ the left-censored fraction or 'clear' fraction.

Let us say our observations $\mathcal{N}$ are sourced from a 'truth' triangle $p_\triangle(S;\boldsymbol{\alpha}^t)$, where the superscript 't' is short for 'truth'. We seek an analytic maximum *a posteriori* (MAP) solution $\hat{\boldsymbol{\alpha}}$ of $\nabla_{\boldsymbol{\alpha}}\ln p_{\boldsymbol{A}}(\boldsymbol{\alpha}|\mathcal{N})=\boldsymbol{0}$. Our goal is to see how close $\hat{\boldsymbol{\alpha}}$ is to the $\boldsymbol{\alpha}^t$ of the underlying truth.

We will consider the case $N\to\infty$ of a perfectly sampled truth triangle. This allows us to convert the above sum in (D5) to an integral, yielding

$$\begin{aligned}J_\infty \equiv \lim_{N\to\infty}N^{-1}\ln p_{\boldsymbol{A}}(\boldsymbol{\alpha}|\boldsymbol{\alpha}^t) &= f'(\boldsymbol{\alpha}^t)\ln f'(\boldsymbol{\alpha})\\ &+\int_1^\infty p_S(s;\boldsymbol{\alpha}^t)\ln p_S(s;\boldsymbol{\alpha})\,\mathrm{d}s,\end{aligned} \tag{D8}$$

where we have dropped the prior term, since it is rendered insignificant in the presence of a perfectly sampled truth. The MAP solution $\hat{\boldsymbol{\alpha}}$ is the $\boldsymbol{\alpha}$ that maximizes J_∞.

D1. Truth is clear

If the truth is wholly clear, i.e. $f'(\boldsymbol{\alpha}^t)=1$, then the truth triangle has no overlap with $s\ge 1$ and so the integral term is zero. This leaves $J_\infty=\ln f'(\boldsymbol{\alpha})$, which is maximized at zero (i.e. $f'(\boldsymbol{\alpha})=1$) for *any* triangle $\boldsymbol{\alpha}$ that is also wholly clear. That is the limit of the information available from the observations in this case – namely that the truth is wholly clear.

D2. Truth is partially cloudy with $S_*^t\le 1$

Next, say the truth is partially cloudy, with $f'(\boldsymbol{\alpha}^t)\in[P_\mathrm{L}^t,1)\subset(0,1)$, which means that $S_*^t\le 1$ and the cloudy portion of the truth triangle is wholly within the falling upper section of the triangle. Then

$$\begin{aligned}J_\infty &= [1-(S_\mathrm{H}^t-1)^2/\theta_\mathrm{H}^t]\ln f'(\boldsymbol{\alpha})\\ &+\frac{2}{\theta_\mathrm{H}^t}\int_1^{S_\mathrm{H}^t}(S_\mathrm{H}^t-s)\ln p_S(s;\boldsymbol{\alpha})\,\mathrm{d}s.\end{aligned} \tag{D9}$$

Clearly, to avoid $J_\infty=-\infty$, i.e. zero *a posteriori* probability, we require $f'(\boldsymbol{\alpha})>0$ ($\boldsymbol{\alpha}$ not overcast) and $p_S(s;\boldsymbol{\alpha})>0$ on $s\in(1,S_\mathrm{H}^t)$, so that $S_\mathrm{H}\ge S_\mathrm{H}^t>1$ (which also prohibits $f'(\boldsymbol{\alpha})=1$, the clear case).

D2.1. $S_*\le 1$

First, consider the case where $S_*\le 1$ also, so that $f'(\boldsymbol{\alpha})\in[P_\mathrm{L},1)\subset(0,1)$ and the cloudy portion is also in the upper section of triangle $\boldsymbol{\alpha}$. Then

$$\begin{aligned}J_\infty &= \left[1-\frac{(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}^t}\right]\ln\left[1-\frac{(S_\mathrm{H}-1)^2}{\theta_\mathrm{H}}\right]\\ &+\int_1^{S_\mathrm{H}^t}\left[\frac{2\,(S_\mathrm{H}^t-s)}{\theta_\mathrm{H}^t}\right]\ln\left[\frac{2\,(S_\mathrm{H}-s)}{\theta_\mathrm{H}}\right]\mathrm{d}s.\end{aligned} \tag{D10}$$

Immediately, we can see there is also an *identifiability problem* in this case, because this J_∞ is parametrized by only two (non-truth) parameters, S_H and $\theta_\mathrm{H}=P_\mathrm{H}\Delta^2$, rather than three. This means that any $\boldsymbol{\alpha}$ that has the same S_H and θ_H also has the same J_∞. This means that there is a family of points in $\boldsymbol{\alpha}$ space, a contour surface in J_∞, given by S_H and $P_\mathrm{H}\Delta^2$ equal to constants. This not only applies in general, but also to the specific values of S_H and θ_H that maximize J_∞. In other words, the MAP solution $\hat{\boldsymbol{\alpha}}$ is not unique but refers to the contour surface with this maximum J_∞, as above. The cause of this problem is that the PDF is purely linear for this case, thereby involving only two rather than three independent parameters. In retrospect, it may have been better to have used a smooth three-parameter PDF, such as the GEV distribution used in Norris *et al.* (2008).

We proceed to evaluate (D10) and its MAP solution. We do so on the domain $S_\mathrm{H}\ge S_\mathrm{H}^t>1$, as required earlier for $J_\infty>-\infty$. The condition $S_*\le 1$, i.e. $f'(\boldsymbol{\alpha})\in[P_\mathrm{L},1)\subset(0,1)$, ensures $f'(\boldsymbol{\alpha})>0$, also required for $J_\infty>-\infty$. After some algebra,

$$\begin{aligned}J_\infty &= \left[1-\frac{(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}^t}\right]\ln\left[1-\frac{(S_\mathrm{H}-1)^2}{\theta_\mathrm{H}}\right]\\ &+\frac{(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}^t}\ln\frac{2}{\theta_\mathrm{H}}+\frac{1}{2\theta_\mathrm{H}^t}\\ &\times\Big\{(S_\mathrm{H}-1)\,[4(S_\mathrm{H}-S_\mathrm{H}^t)-(S_\mathrm{H}-1)\\ &+2((S_\mathrm{H}-1)-2(S_\mathrm{H}-S_\mathrm{H}^t))\ln(S_\mathrm{H}-1)]\\ &-(S_\mathrm{H}-S_\mathrm{H}^t)^2\,[3-2\ln(S_\mathrm{H}-S_\mathrm{H}^t)]\Big\}.\end{aligned} \tag{D11}$$

As discussed earlier, we may regard J_∞ as a function of S_H and θ_H. The partial derivatives are as follows:

$$\begin{aligned}\frac{\partial J_\infty}{\partial\theta_\mathrm{H}} &= \frac{(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}^t\theta_\mathrm{H}}\\ &\times\left[\frac{\theta_\mathrm{H}^t-(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}-(S_\mathrm{H}-1)^2}\cdot\frac{(S_\mathrm{H}-1)^2}{(S_\mathrm{H}^t-1)^2}-1\right]\end{aligned} \tag{D12}$$

and

$$\frac{\partial J_\infty}{\partial S_\mathrm{H}} = -\frac{2(S_\mathrm{H}-1)}{\theta_\mathrm{H}^t}\left[\frac{\theta_\mathrm{H}^t-(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}-(S_\mathrm{H}-1)^2} - 1 \right. \\ \left. - \frac{S_\mathrm{H}-S_\mathrm{H}^t}{S_\mathrm{H}-1}\left(\ln\frac{S_\mathrm{H}-S_\mathrm{H}^t}{S_\mathrm{H}-1}-1\right)\right]. \tag{D13}$$

Zeroing the partial derivative with respect to θ_H yields

$$\frac{\theta_\mathrm{H}^t}{\theta_\mathrm{H}} = \left(\frac{S_\mathrm{H}^t-1}{S_\mathrm{H}-1}\right)^2. \tag{D14}$$

Similarly, zeroing the partial derivative with respect to S_H yields

$$\frac{\theta_\mathrm{H}^t-(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}-(S_\mathrm{H}-1)^2} - 1 \\ = \frac{S_\mathrm{H}-S_\mathrm{H}^t}{S_\mathrm{H}-1}\left(\ln\frac{S_\mathrm{H}-S_\mathrm{H}^t}{S_\mathrm{H}-1}-1\right). \tag{D15}$$

One obvious solution of these equations is $S_\mathrm{H} = S_\mathrm{H}^t$ and $\theta_\mathrm{H} = \theta_\mathrm{H}^t$ (note that the right-hand side of the last equation goes to zero as $S_\mathrm{H} \to S_\mathrm{H}^t$). Are there any other solutions? Well, substituting (D14) into (D15) and assuming $S_\mathrm{H} \neq S_\mathrm{H}^t$ yields

$$y \equiv x + \ln(1-x) = 0, \quad x = \frac{S_\mathrm{H}^t-1}{S_\mathrm{H}-1} \in (0,1). \tag{D16}$$

It is simple to show this has no solution on the required $x \in (0,1)$. Thus, the only critical points of J_∞ are on the boundary $S_\mathrm{H} = S_\mathrm{H}^t$. On this boundary, it is easy to verify that not only is $\theta_\mathrm{H} = \theta_\mathrm{H}^t$ the sole critical point, but it is in fact a global maximum on $S_\mathrm{H} = S_\mathrm{H}^t$.

There is an intricacy we have thus far ignored: (D10) and the following equations are only valid for $S_* \leq 1$. This puts a lower limit on θ_H for a given S_H, but a limit that cannot be delineated in terms of S_H only; rather, it requires another parameter, let us say P_H. Specifically, $S_* \leq 1 \implies S_\mathrm{H} - \sqrt{P_\mathrm{H}\theta_\mathrm{H}} \leq 1 \implies \theta_\mathrm{H} \geq (S_\mathrm{H}-1)^2/P_\mathrm{H} \equiv \theta_\mathrm{H}^\mathrm{min}(S_\mathrm{H}, P_\mathrm{H})$. In addition, we also require (on physical grounds) that $S_\mathrm{L} \geq 0 \implies \theta_\mathrm{H} \leq P_\mathrm{H}S_\mathrm{H}^2 \equiv \theta_\mathrm{H}^\mathrm{max}(S_\mathrm{H}, P_\mathrm{H})$. (Note that $\theta_\mathrm{H}^\mathrm{min} < \theta_\mathrm{H}^\mathrm{max} \iff S_\mathrm{H} < 1/(1-P_\mathrm{H}) \in [1.11, 10]$, since, in practice, we use $P_\mathrm{H} \in [0.1, 0.9] \equiv [P_\mathrm{H}^\mathrm{min}, P_\mathrm{H}^\mathrm{max}]$. However, since we also use $S_\mathrm{H} \leq S_\mathrm{max} = 1.1$, at least for water clouds, then $[\theta_\mathrm{H}^\mathrm{min}, \theta_\mathrm{H}^\mathrm{max}]$ is a valid interval without any further restrictions. We will not discuss the ice case here.)

Therefore, we can actually think of J_∞ as a function on a closed bounded domain in the three-dimensional $(S_\mathrm{H}, \theta_\mathrm{H}, P_\mathrm{H})$ space. While $\partial J_\infty/\partial P_\mathrm{H} = 0$, the P_H is important in setting the domain, namely $\theta_\mathrm{H} \in [\theta_\mathrm{H}^\mathrm{min}, \theta_\mathrm{H}^\mathrm{max}]$, as above. The domain in the other dimensions is $S_\mathrm{H} \in [S_\mathrm{H}^t, S_\mathrm{max}]$ and $P_\mathrm{H} \in [P_\mathrm{H}^\mathrm{min}, P_\mathrm{H}^\mathrm{max}]$.

Now, what our earlier analysis showed was firstly that, for $S_\mathrm{H} > S_\mathrm{H}^t$, there are no critical points and hence no extrema of J_∞ in the *interior* of the above domain. Thus, the global maximum of J_∞ must lie on the boundaries of the domain. Secondly, we showed that there was a global maximum on the boundary $S_\mathrm{H} = S_\mathrm{H}^t$ at $\theta_\mathrm{H} = \theta_\mathrm{H}^t$. This is a candidate for the global maximum of the whole domain, if indeed $\theta_\mathrm{H}^t \in [\theta_\mathrm{H}^\mathrm{min}, \theta_\mathrm{H}^\mathrm{max}]$ for $S_\mathrm{H} = S_\mathrm{H}^t$ and for at least some $P_\mathrm{H} \in [P_\mathrm{H}^\mathrm{min}, P_\mathrm{H}^\mathrm{max}]$. We can be sure this is at least true for $P_\mathrm{H} = P_\mathrm{H}^t$, since then the test triangle is identical with the truth triangle, which is certainly in the domain. To find what other P_H are acceptable, we must evaluate $(S_\mathrm{H}^t-1)^2/P_\mathrm{H} \leq \theta_\mathrm{H}^t \leq P_\mathrm{H}S_\mathrm{H}^{t2} \iff P_\mathrm{H} \geq P_\mathrm{H}^*$, where

$$P_\mathrm{H}^* \equiv \max[(S_\mathrm{H}^t-1)^2/\theta_\mathrm{H}^t, \theta_\mathrm{H}^t/S_\mathrm{H}^{t2}]. \tag{D17}$$

Hence, $P_\mathrm{H} \in [\max(P_\mathrm{H}^\mathrm{min}, P_\mathrm{H}^*), P_\mathrm{H}^\mathrm{max}]$ is the range for which $S_\mathrm{H} = S_\mathrm{H}^t$ and $\theta_\mathrm{H} = \theta_\mathrm{H}^t$ is a possible global maximum. (Again, we do not have to worry that the above range is empty, i.e. that $P_\mathrm{H}^* > P_\mathrm{H}^\mathrm{max}$, since know that at least P_H^t is an included point.)

Actually, to be sure that $S_\mathrm{H} = S_\mathrm{H}^t$ and $\theta_\mathrm{H} = \theta_\mathrm{H}^t$ is *the* global maximum, for the above P_H range, we must see if any other point on the other boundaries exceeds its J_∞ value. We will not do this for two reasons: firstly it will be tedious and secondly it seems eminently reasonable to us that the maximum J_∞ for infinite observations should be attainted when the test triangle is in fact the same as the truth triangle.

So, *in conclusion*, the MAP solution for this particular case (namely a partially cloudy truth with $S_*^t \leq 1$ and testing for $\hat{S}_* \leq 1$ also) is given by

$$\hat{S}_\mathrm{H} = S_\mathrm{H}^t \quad \text{and} \quad \hat{\theta}_\mathrm{H} = \theta_\mathrm{H}^t, \tag{D18}$$

and $\hat{P}_\mathrm{H} \in [\max(P_\mathrm{H}^\mathrm{min}, P_\mathrm{H}^*), P_\mathrm{H}^\mathrm{max}]$, an interval that includes P_H^t. This means that the MAP solution is not a unique triangle, but the family of triangles with the same upper bound $\hat{S}_\mathrm{H}$ as the truth, with $\hat{P}_\mathrm{H}\hat{\Delta}^2$ equal to that of the truth and with $\hat{P}_\mathrm{H}$ equal to any value in the range above.

The actual value of J_∞ at this solution is

$$\hat{J}_\infty = \left[1 - \frac{(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}^t}\right]\ln\left[1 - \frac{(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}^t}\right] \\ + \frac{(S_\mathrm{H}^t-1)^2}{\theta_\mathrm{H}^t}\left[\ln\frac{2(S_\mathrm{H}^t-1)}{\theta_\mathrm{H}^t} - \frac{1}{2}\right]. \tag{D19}$$

D2.2. $S_* > 1$

This case, where the test triangle has some cloud in the rising section as well, does not yield a MAP solution. Please contact the authors for the calculations, if required.

D3. Conclusions

We will terminate our analysis here. We will not examine the more complex case where $S_*^t > 1$. In this case, we expect that an identifiability problem will not occur, since both sides of the triangle and therefore all three parameters come into play. *Rather, it is enough for us to realize that a unique MAP solution is not possible for truth triangles that have cloud only in the falling upper leg of the PDF.*

References

Andrieu C, de Freitas N, Doucet A, Jordan MI. 2003. An introduction to MCMC for machine learning. *Mach. Learn.* **50**: 5–43.

Bauer P, Auligńé T, Bell W, Geer A, Guidard V, Heilliette S, Kazumori M, Kim M-J, Liu EH-C, McNally AP, Macpherson B, Okamoto K, Renshaw R, Riishøjgaard L-P. 2011. Satellite cloud and precipitation assimilation at operational NWP centers. *Q. J. R. Meteorol. Soc.* **137**: 1934–1951.

Bodas-Salcedo A, Webb MJ, Bony S, Chepfer H, Dufresne J-L, Klein SA, Zhang Y, Marchand R, Haynes JM, Pincus R, John VO. 2011. COSP: satellite simulation software for model assessment. *Bull. Am. Meteorol. Soc.* **92**: 1023–1043, doi: 10.1175/2011BAMS2856.1.

Cowles MK, Carlin BP. 1996. Markov Chain Monte Carlo convergence diagnostics: A comparative review. *J. Am. Stat. Assoc.* **91**: 883–904.

Daley R. 1991. *Atmospheric Data Analysis*. Cambridge University Press: Cambridge, UK.

Dee DP, da Silva AM. 2003. The choice of variable for atmospheric moisture analysis. *Mon. Weather Rev.* **131**: 155–171.

Gaspari G, Cohn SE, Guo J, Pawson S. 2006. Construction and application of covariance functions with variable length-fields. *Q. J. R. Meteorol. Soc.* **132**: 1815–1838.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. *Bayesian Data Analysis* (2nd edn). Chapman & Hall: New York, NY.

Klein SA, Jakob C. 1999. Validation and sensitivities of frontal clouds simulated by the ECMWF model. *Mon. Weather Rev.* **127**: 2514–2531.

Liu JS, Liang F, Wong WH. 2000. The use of multiple-try method and local optimization in Metropolis sampling. *J. Am. Stat. Assoc.* **95**: 121–134.

Noh YJ, Seaman CJ, Vonder Haar TH, Liu G. 2013. In situ aircraft measurements of the vertical distribution of liquid and ice water content in midlatitude mixed-phase clouds. *J. Appl. Meteorol. Clim.* **52**: 269–279.

Norris PM, da Silva AM. 2007. Assimilation of satellite cloud data into the GMAO finite volume data assimilation system using a parameter estimation method. *J. Atmos. Sci.* **64**: 3880–3895.

Norris PM, Oreopoulos L, Hou AY, Tao WK, Zeng X. 2008. Representation of 3D heterogeneous cloud fields using copulas: Theory for water clouds. *Q. J. R. Meteorol. Soc.* **134**: 1843–1864, doi: 10.1002/qj.321.

Oreopoulos L, Norris PM. 2011. An analysis of cloud overlap at a midlatitude atmospheric observation facility. *Atmos. Chem. Phys.* **11**: 5557–5567, doi: 10.5194/acp-11-5557-2011.

Pincus R, Barker HW, Morcrette J-J. 2003. A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous cloud fields. *J. Geophys. Res.* **108**(D13): 4376, doi: 10.1029/2002JD003322.

Pincus R, Platnick S, Ackerman SA, Hemler RS, Hofmann RJP. 2012. Reconciling simulated and observed views of clouds: MODIS, ISCCP, and the limits of instrument simulators. *J. Clim.* **25**: 4699–4720.

Posselt DJ. 2013. Markov Chain Monte Carlo methods: Theory and applications. In *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications* (Vol. II), Park SK, Xu L. (eds.): 59–87. Springer-Verlag Berlin Heidelberg.

Roberts GO, Rosenthal JS. 2001. Optimal scaling for various Metropolis–Hastings algorithms. *Stat. Sci.* **16**: 351–367.

Tompkins AM. 2002. A prognostic parameterization for the subgrid-scale variability of water vapor and clouds in large-scale models and its use to diagnose cloud cover. *J. Atmos. Sci.* **59**: 1917–1942.

Wind G, Platnick S, King MD, Hubanks PA, Pavolonis MJ, Heidinger AK, Yang P, Baum BA. 2010. Multilayer cloud detection with the MODIS near-infrared water vapor absorption band. *J. Appl. Meteorol. Climatol.* **49**: 2315–2333.

Wind G, da Silva AM, Norris PM, Platnick S. 2013. Multi-sensor cloud retrieval simulator and remote sensing from model parameters – Part 1: Synthetic sensor radiance formulation. *Geosci. Model Dev.* **6**: 2049–2062, doi: 10.5194/gmd-6-2049-2013.

Yang PLZ, Hong G, Nasiri SL, Baum BA, Huang HL, King MD, Platnick S. 2007. Differences between collection 004 and 005 MODIS ice cloud optical/microphysical products and their impact on radiative forcing simulations. *IEEE Trans. Geosci. Remote Sens.* **45**: 2886–2899.

Zhang Z, Platnick S. 2011. An assessment of differences between cloud effective particle radius for marine water clouds from three MODIS spectral bands. *J. Geophys. Res.* **116**: D20215, doi: 10.1029/2011JD016216.